# Model Selection for Mixture Models-Perspectives and Strategies

Gilles Celeux, Sylvia Frühwirth-Schnatter, Christian Robert

# 7

# *Model Selection for Mixture Models – Perspectives and Strategies*

**Gilles Celeux, Sylvia Frühwirth-Schnatter and Christian P. Robert**

*INRIA Saclay, France; Vienna University of Economics and Business, Austria; Université Paris-Dauphine, France and University of Warwick, UK*

**CONTENTS**

## 7.1 Introduction

Determining the number $G$ of components in a finite mixture distribution defined as

$$y \sim \sum_{g=1}^{G} \eta_g f_g(y|\theta_g), \tag{7.1}$$

is an important and difficult issue. This is a most important question, because statistical inference about the resulting model is highly sensitive to the value of $G$. Selecting an erroneous value of $G$ may produce a poor density estimate. This is also a most difficult question from a theoretical perspective as it relates to unidentifiability issues of the mixture model, as discussed already in Chapter 4. This is a most relevant question from a practical viewpoint since the meaning of the number of components $G$ is strongly related to the modelling purpose of a mixture distribution.

From this perspective, the famous quote from Box (1976), "All models are wrong, but some are useful" is particularly relevant for mixture models since they may be viewed as a semi-parametric tool when addressing the general purpose of density estimation or as a model-based clustering tool when concerned with unsupervised classification; see also Chapter 1. Thus, it is highly desirable and ultimately profitable to take into account the grand modelling purpose of the statistical analysis when selecting a proper value of $G$, and we distinguish in this chapter between selecting $G$ as a density estimation problem in Section 7.2 and selecting $G$ in a model-based clustering framework in Section 7.3.

Both sections will discuss frequentist as well as Bayesian approaches. At a foundational level, the Bayesian approach is often characterized as being highly directive, once the prior distribution has been chosen (see, for example, Robert, 2007). While the impact of the prior on the evaluation of the number of components in a mixture model or of the number of clusters in a sample from a mixture distribution cannot be denied, there exist competing ways of assessing these quantities, some borrowing from point estimation and others from hypothesis testing or model choice, which implies that the solution produced will strongly depend on the perspective adopted. We present here some of the Bayesian solutions to the different interpretations of picking the "right" number of components in a mixture, before concluding on the ill-posed nature of the question.

As already mentioned in Chapter 1, there exists an intrinsic and foundational difference between frequentist and Bayesian inferences: only Bayesians can truly *estimate* $G$, that is, treat $G$ as an additional unknown parameter that can be estimated simultaneously with the other model parameters $\theta = (\eta_1, \ldots, \eta_G, \theta_1, \ldots, \theta_G)$ defining the mixture distribution (7.1). Nevertheless, Bayesians very often rely on model selection perspectives for $G$, meaning that Bayesian inference is carried out for a range of values of $G$, from 1, say, to a pre-specified maximum value $G_{\max}$, given a sample $\mathbf{y} = (y_1, \ldots, y_n)$ from (7.1). Each value of $G$ thus corresponds to a potential model $\mathcal{M}_G$, and those models are compared via Bayesian model selection. A typical choice for conducting this comparison is through the values of the marginal likelihood $p(\mathbf{y}|G)$,

$$p(\mathbf{y}|G) = \int p(\mathbf{y}|\theta, G)p(\theta|G)d\theta, \tag{7.2}$$

separately for each mixture model $\mathcal{M}_G$, with $p(\theta|G)$ being a prior distribution for all unknown parameters $\theta$ in a mixture model with $G$ components.

However, cross-model Bayesian inference on $G$ is far more attractive, at least conceptually, as it relies on one-sweep algorithms, namely computational procedures that yield

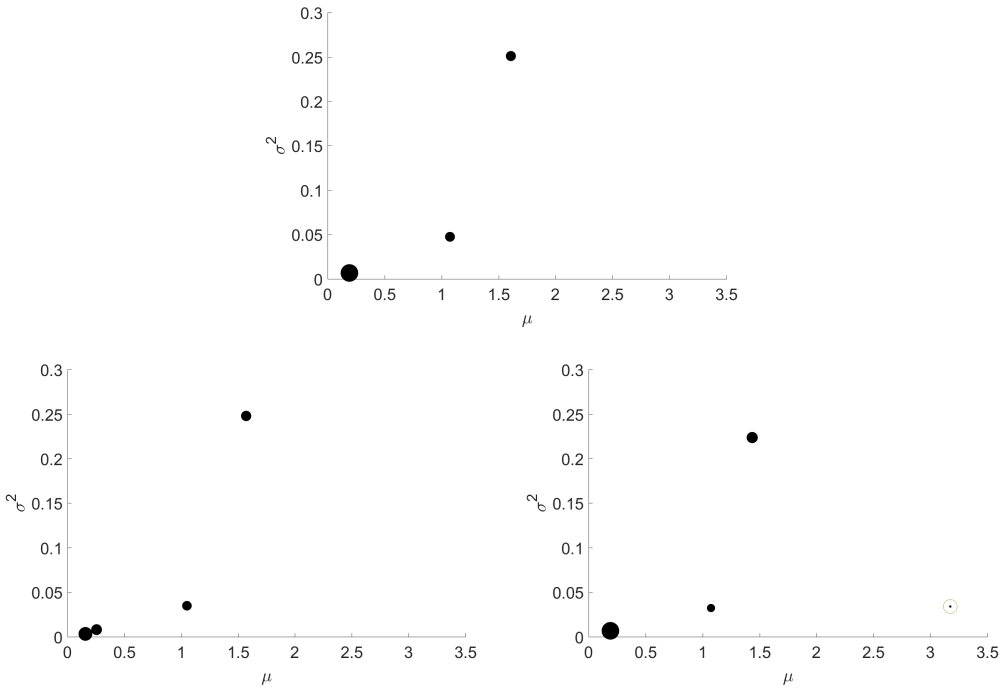


**FIGURE 7.1**
Point process representation of the estimated mixture parameters for three mixture distributions fitted to the enzyme data using a Bayesian framework under the prior of Richardson & Green (1997). The size of each point $(\hat{\mu}_g, \hat{\sigma}_g^2)$ corresponds to the mixture weight $\hat{\eta}_g$. Top: $G = 3$. Bottom: $G = 4$ with $\eta \sim \mathcal{D}_4(4)$ (left) and $\eta \sim \mathcal{D}_4(0.5)$ (right; the very small fourth component is marked by a circle).

estimators of $G$ jointly with the unknown model parameters. Section 7.4 reviews such one-sweep Bayesian methods for cross-model inference on $G$, ranging from well-known methods such as reversible jump Markov chain Monte Carlo (MCMC) to more recent ideas involving sparse finite mixtures relying on overfitting in combination with a prior on the weight distribution that forces sparsity.

## 7.2 Selecting $G$ as a Density Estimation Problem

When the estimation of the data distribution is the main purpose of the mixture modelling, it is generally assumed that this distribution truly is a finite mixture distribution. One inference issue is then to find the true number of mixture components, $G$, that is, the *order* of the mixture behind the observations. This assumption is supposed to produce well-grounded tests and model selection criteria.

The *true order* of a finite mixture model is the smallest value of $G$ such that the components of the mixture in (7.1) are all distinct and the mixing proportions are all positive (that is, $\theta_g \neq \theta_{g'}$, $g \neq g'$ and $\eta_g > 0$). This definition attempts to deal with the ambiguity (or

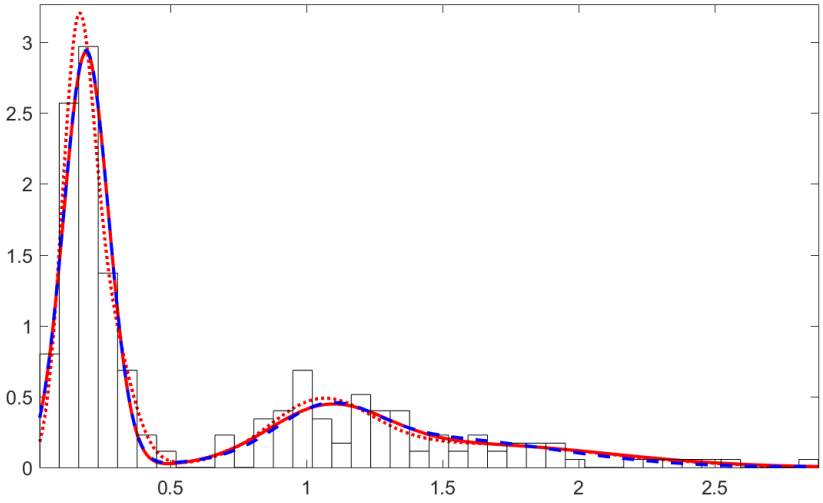


**FIGURE 7.2**
Histogram of the enzyme data together with three fitted mixture distributions: $G = 3$ (solid line); $G = 4$ and $\eta \sim \mathcal{D}_4(4)$ (dotted line); $G = 4$ and $\eta \sim \mathcal{D}_4(0.5)$ (dashed line). The dashed and solid lines are nearly identical.

non-identifiability) due to overfitting, discussed in Section 1.3 of Chapter 1 and Section 4.2.2 of Chapter 4: a mixture with $G$ components can equally be defined as a (non-identifiable) mixture with $G + 1$ components where the additional component either has a mixing proportion $\eta_{G+1}$ equal to zero or the parameter $\theta_{G+1}$ is identical to the parameter $\theta_g$ of some other component $g \in \{1, \ldots, G\}$. These identifiability issues impact both frequentist and Bayesian methods for selecting $G$. Hence, the order $G$ is a poorly defined quantity and in practical mixture analysis it is often difficult to decide what order $G$ describes the data best.

By way of illustration, a mixture of normal distributions $\mathcal{N}(\mu_g, \sigma_g^2)$ with $G = 3$ components is fitted within a Bayesian framework to the enzyme data studied in Richardson & Green (1997), using the same prior as Richardson & Green, in particular a uniform prior on the weight distribution $\eta = (\eta_1, \ldots, \eta_G)$. In addition, mixtures with $G = 4$ components are fitted, but with different symmetric Dirichlet priors for $\eta$, namely $\eta \sim \mathcal{D}_4(4)$ and $\eta \sim \mathcal{D}_4(0.5)$. As discussed in Section 4.2.2 above, the first prior favours overlapping components, whereas the second prior favours small components, should the mixture be overfitting.

Full conditional Gibbs sampling is applied for posterior inference. All three mixture models are identified by $k$-means clustering in the point process representation of the posterior draws of $(\mu_g, \sigma_g)$. The estimated component parameters $(\hat{\mu}_g, \hat{\sigma}_g^2, \hat{\eta}_g)$ are visualized through a point process representation in Figure 7.1. Obviously, the parameters for the four-component mixture are quite different and emerge in quite different ways than the components of the three-component mixture. The component $(\hat{\mu}_g, \hat{\sigma}_g^2, \eta_g) = (0.19, 0.007, 0.61)$ is split into the two components $(\hat{\mu}_g, \hat{\sigma}_g^2) = (0.16, 0.003)$ and $(\hat{\mu}_{g'}, \hat{\sigma}_{g'}^2) = (0.26, 0.008)$ with weights $0.38 + 0.23 = 0.61$ under the prior $\eta \sim \mathcal{D}_4(4)$. Under the prior $\eta \sim \mathcal{D}_4(0.5)$, the vari-

ance of the two components with the larger means is reduced and a fourth tiny component with weight 0.012 and a large mean are added.

Figure 7.2 shows the density of these three mixture distributions together with a histogram of the data. The density of $G = 4$ under the prior $\eta \sim \mathcal{D}_4(0.5)$ is nearly identical to the density of $G = 3$ with the tiny fourth component capturing the largest observations. The density of $G = 4$ under the prior $\eta \sim \mathcal{D}_4(4)$ is also very similar to the density of $G = 3$, but tries to capture the skewness in the large, well-separated cluster with the smallest observations. Clearly, it is not easy to decide which of these three densities describes the data best.

### 7.2.1 Testing the order of a finite mixture through likelihood ratio tests

From a frequentist perspective, a natural approach to the determination of the order of a mixture distribution is to rely on the likelihood ratio test associated with the hypotheses of $G$ ($H_0$) versus $G+1$ ($H_A$) non-empty components. However, as a consequence of the above-mentioned identifiability problem, regularity conditions ensuring a standard asymptotic distribution for the maximum likelihood (ML) estimates do not hold; see Section 4.3.3. When one component is superfluous ($H_0$), the parameter $\theta_{G+1}$ under the alternative hypothesis $H_A$ lies on the boundary of the parameter space. Moreover, the remainder term appearing within a series expansion of the likelihood ratio test statistic is not uniformly bounded under $H_A$. Therefore, its distribution remains unknown.

Many attempts have been made to modify the likelihood ratio test in this setting; see, for example, the references in McLachlan & Peel (2000) and Frühwirth-Schnatter (2006). Here, we wish to mention the seminal works of Dacunha-Castelle & Gassiat (1997, 1999), which make use of a locally conic parameterization to deal with non-identifiability. This research has been updated and extended to ensure a consistent estimation of $G$ with penalized ML when $G$ is bounded for independent and dependent finite mixtures (Gassiat, 2002). Note that this boundary on $G$ has been relaxed in the paper of Gassiat & van Handel (2013) for a mixture of translated distributions. Moreover, an early reference that deals explicitly with testing $G$ against $G + 1$ in Markov switching models (see Chapter 13) is Hansen (1992).

Adopting a different perspective, McLachlan (1987) proposed using a parametric bootstrap test to select the number of components in a normal mixture. This approach can be extended without difficulty to other mixture distributions. To test the null hypothesis that $G = G_0$ against the alternative that $G = G_1$ at the level $\alpha$, McLachlan (1987) suggests the following procedure: draw $B$ bootstrap samples from a mixture model of order $G_0$ with the parameters being equal to the maximum likelihood estimator (MLE) $\hat{\theta}_{G_0}$ and compute the log likelihood ratio statistic (LRS) of $G = G_0$ versus $G = G_1$ for all bootstrap samples. If the LRS computed on the original sample is smaller than the $1 - \alpha$ quantile of the distribution of the bootstrapped LRSs, then the hypothesis $G = G_0$ is not rejected. It must be pointed out that this bootstrap test is *biased* since the $p$-value is computed from a bootstrap sample where the parameter value $\theta_{G_0}$ has been estimated from the *whole* observed sample. One way to address this bias is to resort to *double bootstrapping*: first, $B$ bootstrap samples are used to compute an estimate $\hat{\theta}^b_{G_0}$ for each bootstrap sample $b = 1, \ldots, B$, while a second bootstrap layer produces an LRS for each bootstrap sample $b$ of the first bootstrap layer. Unfortunately, this double bootstrap procedure is extremely computer-intensive.

As far as we know, technical difficulties aside, statistical tests are rarely used to estimate the order of a mixture. There are several reasons for this. First, the mixture models under comparison are not necessarily embedded. And second, the proposed tests are numerically difficult to implement and slow. Hence, other procedures such as optimizing penalized log likelihood or resorting to Bayesian methods are preferable.

### 7.2.2 Information criteria for order selection

Various information criteria for selecting the order of a mixture distribution are discussed in this section, including the Akaike (AIC) and Bayesian (BIC) information criteria (Section 7.2.2.1), the slope heuristic (Section 7.2.2.2), the deviance information criterion (DIC) (Section 7.2.2.3) , and the minimum message length (Section 7.2.2.4) and we refer to the literature for additional criteria such as the approximate weight of evidence (AWE) criterion (Banfield & Raftery, 1993). Information criteria are based on penalizing the log of the likelihood function $\mathrm{L}_o(\theta; G)$, also known as the observed-data likelihood,

$$\mathrm{L}_o(\theta; G) = \prod_{i=1}^{n} \left[ \sum_{g=1}^{G} \eta_g f_g(y_i \mid \theta_g) \right], \tag{7.3}$$

of a mixture model $\mathcal{M}_G$ with $G$ components. The penalty is proportional to the number of free parameters in $\mathcal{M}_G$, denoted by $\upsilon_G$, and the various criteria differ in the choice of the corresponding proportionality factor. The number $\upsilon_G$ increases linearly in $G$ and quantifies the complexity of the model. For a multivariate mixture of Gaussian distributions with unconstrained covariance matrices generating observations of dimension $r$, for instance, $\upsilon_G = G(1 + r + r(r+1)/2) - 1$.

#### 7.2.2.1 AIC and BIC

Let $\hat{\theta}_G$ be the MLE corresponding to the observed-data likelihood $\mathrm{L}_o(\theta; G)$, defined in (7.3). The AIC (Akaike, 1974) and BIC (Schwarz, 1978) are popular model selection criteria for solving the bias–variance dilemma for choosing a parsimonious model. $\mathrm{AIC}(G)$ is defined as

$$\mathrm{AIC}(G) = -2 \log \ell_o(\hat{\theta}_G; G) + 2\upsilon_G, \tag{7.4}$$

whereas $\mathrm{BIC}(G)$ is defined as

$$\mathrm{BIC}(G) = -2 \log \ell_o(\hat{\theta}_G; G) + \upsilon_G \log(n). \tag{7.5}$$

Both criteria are asymptotic criteria and assume that the sampling pdf is within the model collection. On the one hand, the AIC aims to minimize the Kullback–Leibler divergence between model $\mathcal{M}_G$ and the sampling pdf. On the other hand, the BIC approximates the marginal likelihood of model $\mathcal{M}_G$, defined in (7.2), by ignoring the impact of the prior.

In some settings and under proper regularity conditions, the BIC can be shown to be consistent, meaning it eventually picks the true order of the mixture, while the AIC is expected to have a good predictive behaviour and happens to be minimax optimal, that is, to minimize the maximum risk among all estimators, in some regular situations (Yang, 2005). However, in a mixture setting both penalized log likelihood criteria face the same difficulties as the likelihood ratio test due to the identifiability problems mentioned in Section 7.2.1.

When regularity conditions hold, the BIC is derived by a Laplace approximation of the marginal likelihood of model $\mathcal{M}_G$:

$$p(\mathbf{y}|G) = \int_{R^{\upsilon_G}} \exp(nL(\theta)) \mathrm{d}\theta = \exp\left(nL(\theta_G^*)\right) \frac{2\pi}{n}^{\upsilon_G/2} \left|-\mathrm{L}''(\theta_G^*)\right|^{-1/2} + O(n^{-1}),$$

where $L : R^{\upsilon_G} \longrightarrow R$ is a $C^2$ function, here

$$L(\theta) = \frac{1}{n}[\log p(\mathbf{y}|\theta, G) + \log p(\theta|G)],$$

with unique maximum $\theta_G^*$. Moreover, the posterior mode $\theta_G^*$ is approximately equal to

the MLE $\hat{\theta}_G$, and the Hessian of $L$ can be approximated with the inverse of the Fisher information $I(\hat{\theta}_G)^{-1}$. Hence,

$$-2\log p(\mathbf{y}|G) = \mathrm{BIC}(G) + O(1).$$

However, as noted above, the BIC is not truly Bayesian in that it bypasses all terms (first and foremost the prior distribution) that do not depend on $n$. Moreover, if the (data-dependent) prior pdf $p(\theta|G)$ is a normal distribution deduced from the MLE distribution $\mathcal{N}(\hat{\theta}_G, I(\hat{\theta}_G)^{-1})$, then

$$-2\log p(\mathbf{y}|G) = \mathrm{BIC}(G) + O(n^{-1/2}).$$

For this very reason, the BIC is mostly exploited for selecting a model estimated through the ML methodology (and not in a Bayesian way).

Under proper regularity conditions, the BIC enjoys the following asymptotic properties.

(a) The BIC is consistent: if there exists $G^*$ such that the true distribution $p_0$ generating the data is equal to $p(\cdot|G^*)$, then, for $n$ large enough, BIC selects $G^*$.

(b) Even if such a $G^*$ does not exist, good behaviour of the BIC can be expected, if $p_0$ is close to $p(\cdot|G^*)$ for the value $G^*$ selected by the BIC.

Unfortunately, the regularity conditions that validate the above Laplace approximation require the model parameters to be identifiable. As seen above, this is not true in general for most mixture models. However, the BIC has been shown to be consistent when the pdfs of the mixture components are bounded (Keribin, 2002). This is, for example, the case for a Gaussian mixture model with equal covariance matrices. In practice, there is no reason to think that the BIC is not consistent for selecting the number of mixture components when the mixture model is used to estimate a density (see, for instance, Roeder & Wasserman, 1997; Fraley & Raftery, 2002).

For singular models for which the Fisher information matrix is not everywhere invertible, Drton & Plummer (2017) proposed the so-called sBIC criterion. This criterion makes use of the Watanabe (2009) marginal likelihood approximation of a singular model. It is the solution of a fixed point equation approximating the weighted average of the log marginal likelihoods of the models in competition. The sBIC criterion is proven to be consistent. It coincides with the BIC criterion when the model is regular. But, while the usual BIC is in fact not Bayesian, the sBIC is connected to the large-sample behaviour of the log marginal likelihood (Drton & Plummer, 2017).

However, the BIC does not lead to a prediction of the observations that is asymptotically optimal; see Yang (2005) and Drton & Plummer (2017) for further discussion on the comparative properties of the AIC and BIC. In contrast to the BIC criterion, the AIC is known to suffer from a marked tendency to overestimate the true value of $G$ (see, for instance, Celeux & Soromenho (1996) for illustrations). However, a modification of AIC, the so-called AIC3 criterion, proposed in Bozdogan (1987), which replaces the penalty $2\upsilon_G$ with $3\upsilon_G$, provides a good assessment of $G$ when the latent class model is used to estimate the density of categorical data (Nadif & Govaert, 1998). Nevertheless, the theoretical reasons for this interesting behaviour of the AIC3 (in this particular context) remain for the most part mysterious.

Finally, when the BIC is used to select the number of a mixture components for real data, it has a marked tendency to choose a large number of components or even to choose the highest proposed number of components. The reason for this behaviour is once more related to the fact that the penalty of the BIC is independent of the data, apart from the sample size $n$. When the bias in the mixture model does not vanish when the number of components increases, the BIC always increases by adding new mixture components. In a

model-based clustering context, this under-penalization tendency is often counterbalanced by the entropy of the mixture, added to rgw BIC in the ICLbic criterion (see Section 7.3.2.1), which could lead to a compromise between the fit of a mixture model and its ability to produce a sensible clustering of the data. But there are many situations where the entropy of the mixture is not enough for counterbalancing this tendency and, moreover, the ICLbic is not really relevant when the modelling purpose is not related to clustering.

#### 7.2.2.2 The Slope Heuristics

The so-called slope heuristics (Birgé & Massart, 2001, 2007), are a data-driven method to calibrate a penalized criterion that is known up to a multiplicative constant $\kappa$. It has been successfully applied to many situations, and particularly to mixture models when using the observed-data log likelihood; see Baudry et al. (2012). As shown by Baudry (2015), it can be extended without difficulty to other contrasts including the conditional classification log likelihood, which will be defined in Section 7.3.2.1. Roughly speaking, as with the AIC and BIC, the penalty function $\text{pen}(G)$ is assumed to be proportional to the number of free parameters $\upsilon_G$ (i.e. the model dimension), $\text{pen}(G) \propto \kappa\upsilon_G$.

The penalty is calibrated using the data-driven slope estimation (DDSE) procedure, available in the R package `capushe` (Baudry et al., 2012). The method assumes a linear relation between the observed-data log likelihood and the penalty. It is important to note that this assumption must and may easily be verified in practice via a simple plot. Then the DDSE procedure directly estimates the slope of the expected linear relationship between the contrast (here the observed-data log likelihood, but other contrasts such as the conditional classification likelihood are possible) and the model dimension $\upsilon_G$ which is a function of the number $G$ of components. The estimated slope $\kappa$ defines a minimal penalty $\kappa\upsilon_G$ below which smaller penalties give rise to the selection of more complex models, while higher penalties should select models with reasonable complexity. Arguments are provided in Birgé & Massart (2007) and Baudry et al. (2012) that the optimal (oracle) penalty is approximately twice the minimal penalty. Thus, by setting the penalty to be $2\kappa\upsilon_G$, the slope heuristics criterion is defined as

$$\text{SH}(G) = -\log \ell_o(\hat{\theta}_G; G) + 2\kappa\upsilon_G,$$

when considering mixture models in a density estimation framework. For more details about the rationale and the implementation of the slope heuristics, see Baudry et al. (2012).

The slope heuristics method relies on the assumption that the bias of the fitted models decreases as their complexity increases and becomes almost constant for the most complex model. In the mixture model framework, this requires the family of models to be roughly nested. More discussion, technical developments and illustrations are given in Baudry (2015).

The ability of the slope heuristics method, which is not based on asymptotic arguments, to detect the stationarity of the model family bias (namely the fact that the bias becomes almost constant) is of prime relevance. It leads this criterion to propose more parsimonious models than the BIC or even the integrated complete-data likelihood criterion (to be discussed in Section 7.3.2.1). Many illustrations of this practical behaviour can be exhibited in various domains of application of mixture models; see, for instance, a clustering use of the slope heuristics to choose the number of components of a multivariate Poisson mixture with RNASeq transcriptome data (Rau et al., 2015) or in a model-based clustering approach for comparing bike sharing systems (Bouveyron et al., 2015).

#### 7.2.2.3 DIC

In recent years, the deviance information criterion introduced by Spiegelhalter et al. (2002) has become a popular criterion for Bayesian model selection because it is easily computed from posterior draws, using MCMC methods. Like other penalized log likelihood criteria, the DIC involves a trade-off between goodness of fit and model complexity, measured in terms of the so-called effective number of parameters. However, the use of the DIC to choose the order $G$ of a mixture model is not without issues, as discussed by De Iorio & Robert (2002) and Celeux et al. (2006).

To apply the DIC in a mixture context, several decisions have to be made. As for any latent variable model, a first difficulty arises in the choice of the appropriate likelihood function. Should the DIC be based on the *observed-data* log likelihood $\log p(\mathbf{y}|\theta, G)$, the *complete-data* log likelihood $\log p(\mathbf{y}, \mathbf{z}|\theta, G)$ or the *conditional* log likelihood $\log p(\mathbf{y}|\mathbf{z}, \theta, G)$, where $\mathbf{z} = (z_1, \ldots, z_n)$ are the latent allocations generating the data (see also Section 7.3.1)? Second, the calculation of the DIC requires an estimate $\hat{\theta}_G$ of the unknown parameter $\theta$ which may suffer from label switching, making the DIC (which is based on averaging over MCMC draws) unstable. Finally, if the definition of the DIC involves either the complete-data or conditional likelihood, the difficulty that $\mathbf{z}$ is unobserved must be dealt with, either by integrating against the posterior $p(\mathbf{z}|\mathbf{y}, G)$ or by using a plug-in estimator of $\mathbf{z}$ in which case once again the label switching problem must be addressed to avoid instability.

In an attempt to calibrate these difficulties, Celeux et al. (2006) investigate in total eight different DIC criteria. $\mathrm{DIC}_2$, for instance, focuses on the marginal distribution of the data and considers the allocations $\mathbf{z}$ as nuisance parameters. Consequently, it is based on the observed-data likelihood:

$$\mathrm{DIC}_2(G) = -4\mathrm{E}_\theta \left(\log\, p(\mathbf{y}|\theta, G)|\mathbf{y}\right) + 2\log\, p(\mathbf{y}|\hat{\theta}_G, G),$$

where the posterior mode estimator $\hat{\theta}_G$ (which is invariant to label switching) is obtained from the observed-data posterior $p(\theta|\mathbf{y}, G)$ and $\mathrm{E}_\theta$ is the expectation with respect to the posterior $p(\theta|\mathbf{y}, G)$.

Based on several simulation studies, Celeux et al. (2006) recommend using $\mathrm{DIC}_4$ which is based on computing first DIC for the complete-data likelihood function and then integrating over $\mathbf{z}$ with respect to the posterior $p(\mathbf{z}|\mathbf{y}, G)$. This yields

$$\mathrm{DIC}_4(G) = -4\mathrm{E}_{\theta,\mathbf{z}} \left(\log\, p(\mathbf{y}, \mathbf{z}|\theta, G)|\mathbf{y}\right) + 2\mathrm{E}_\mathbf{z} \left(\log\, p(\mathbf{y}, \mathbf{z}|\hat{\theta}_G(\mathbf{z}))|\mathbf{y}\right),$$

where $\hat{\theta}_G(\mathbf{z})$ is the complete-data posterior mode which must be computed for each draw from the posterior $p(\mathbf{z}|\mathbf{y}, G)$. This is straightforward if the complete-data posterior $p(\theta_g|\mathbf{y}, \mathbf{z})$ is available in closed form. If this is not the case, Celeux et al. (2006) instead use the posterior mode estimator $\hat{\theta}_G$ of the observed-data posterior $p(\theta|\mathbf{y})$. This leads to an approximation of $\mathrm{DIC}_4(G)$, called $\mathrm{DIC}_{4a}(G)$, which is shown to be a criterion that penalizes $\mathrm{DIC}_2(G)$ by the expected entropy, defined in (7.16):

$$\mathrm{DIC}_{4a}(G) = \mathrm{DIC}_2(G) + 2\mathrm{E}_\theta \left(\mathrm{ENT}(\theta; G)|\mathbf{y}\right).$$

Both $\mathrm{DIC}_2(G)$ and $\mathrm{DIC}_{4a}(G)$ are easily estimated from (MCMC) draws from the posterior $p(\theta|\mathbf{y}, G)$ by substituting all expectations $\mathrm{E}_\bullet(\cdot|\mathbf{y})$ by an average over the corresponding draws. Note that label switching is not a problem here, because both $\log p(\mathbf{y}|\theta, G)$ and $\mathrm{ENT}(\theta; G)$ are invariant to the labelling of the groups.

However, in practical mixture modelling, the DIC turns out to be very unstable, as shown by Celeux et al. (2006) for the galaxy data (Roeder, 1990). A similar behaviour was observed by Frühwirth-Schnatter & Pyne (2010) who fitted skew-normal mixtures to

Alzheimer disease data under various prior assumptions. While the marginal likelihood selected $G = 2$ with high confidence for all priors, $\mathrm{DIC}_{4a}(G)$ selected $G = 1$, regardless of the chosen prior, whereas the number of components selected by $\mathrm{DIC}_2(G)$ ranged from 2 to 4, depending on the prior.

#### 7.2.2.4 The minimum message length

Assuming that the form of the mixture models is fixed (e.g. Gaussian mixture models with free covariance matrices or Gaussian mixture models with a common covariance matrix), several authors have proposed dealing with the estimation of the mixture parameters and $G$ in a single algorithm with the minimum message length (MML) criterion (see, for instance, Rissanen, 2012; Wallace & Freeman, 1987). Considering the MML criterion in a Bayesian perspective and choosing Jeffreys' non-informative prior $p(\theta)$ for the mixture parameter, Figueiredo & Jain (2002) propose minimizing the criterion

$$\mathrm{MML}(\theta; G) = -\log p(\mathbf{y}|\theta, G) - \log p(\theta|G) + \frac{1}{2}\log|I(\theta)| + \frac{\upsilon_G}{2}(1 - \log(12)),$$

where $I(\theta)$ is the expected Fisher information matrix which is approximated by the complete-data Fisher information matrix $I_C(\theta)$.

As we know, for instance from Section 4.2.2 above, Jeffreys' non-informative prior does not work for mixtures. Figueiredo & Jain (2002) circumvent this difficulty by only considering the parameters of the components whose proportion is non-zero, namely the components $g$ such that $\hat{\eta}_g > 0$.

Assuming, for instance, that the mixture model considered arises from the general Gaussian mixture family with free covariance matrices, this approach leads to minimizing the criterion

$$\begin{aligned}\mathrm{MML}(\theta; G) = &-\log p(\mathbf{y}|\theta, G) + \frac{G^\star}{2}\log\frac{n}{12} \\ &+ \frac{\dim(\theta_g)}{2}\sum_{g:\hat{\eta}_g>0}\{\log(n\cdot\dim(\theta_g)/12) + G^\star(\dim(\theta_g)+1)\},\end{aligned} \tag{7.6}$$

with $G^\star = \mathrm{card}\{g|\hat{\eta}_g > 0\}$. In this Bayesian context, the approach of Figueiredo & Jain (2002) involves optimizing iteratively the criterion (7.6), starting from a large number of components $G_{\max}$, and cancelling the components $g$ such that, at iteration $s$,

$$\sum_{i=1}^{n}\hat{\tau}_{ig}^{(s)} < \frac{\dim(\theta_g^{(s)})}{2}, \tag{7.7}$$

where $\hat{\tau}_{ig}^{(s)}$ are the elements of the fuzzy classification matrix defined in (7.18). Thus, the chosen number of components $G^\star$ is the number of components remaining at the convergence of the iterative algorithm. This iterative algorithm could be the EM algorithm, but Figueiredo & Jain (2002) argue that with EM, for large $G$, it can happen that no component has enough initial support, as the criterion for cancellation defined in (7.7) is fulfilled for *all* $G$ components. Thus, they prefer to make use of the componentwise EM algorithm of Celeux et al. (2001), which updates the $\eta_g$ and the $\theta_g$ sequentially: update $\eta_1$ and $\theta_1$, recompute $\tau_{i1}$ for $i = 1, \ldots, n$, update $\eta_2$ and $\theta_2$, recompute $\tau_{i2}$ for $i = 1, \ldots, n$, and so on.

Zeng & Cheung (2014) use exactly the same approach with the completed-data or the classification likelihood instead of the observed-data likelihood. Thus, roughly speaking, the procedure of Figueiredo & Jain (2002) is expected to provide a similar number of components to the BIC, while the procedure of Zeng & Cheung (2014) is expected to provide a similar number of clusters to the ICLbic presented in Section 7.3.2.1.

### 7.2.3 Bayesian model choice based on marginal likelihoods

From a Bayesian testing perspective, selecting the number of components can be interpreted as a model selection problem, given the probability of each model within a collection of all models corresponding to the different numbers of components (Berger, 1985). The standard Bayesian tool for making this model choice is based on the marginal likelihood (also called *evidence*) of the data $p(\mathbf{y}|G)$ for each model $\mathcal{M}_G$, defined in (7.2), which naturally penalizes models with more components (and more parameters) (Berger & Jefferys, 1992).

While the BIC is often considered as one case of information criterion, it is important to recall (see Section 7.2.2.1) that it was first introduced by Schwartz (1965) as an approximation to the marginal likelihood $p(\mathbf{y}|G)$. Since this approximation does not depend on the choice of the prior $p(\theta|G)$, it is not of direct appeal for a Bayesian evaluation of the number of components, especially when considering that the marginal likelihood itself can be approximated by simulation-based methods, as discussed in this section.

#### 7.2.3.1 Chib's method, limitations and extensions

The reference estimator for evidence approximation is Chib's (1995) representation of the marginal likelihood of model $\mathcal{M}_G$ as[1]

$$p(\mathbf{y}|G) = \frac{p(\mathbf{y}|\theta^o, G)p(\theta^o|G)}{p(\theta^o|\mathbf{y}, G)}, \tag{7.8}$$

which holds for *any* choice of the plug-in value $\theta^o$. While the posterior $p(\theta^o|\mathbf{y}, G)$ is not available in closed form for mixtures, a Gibbs sampling decomposition allows for a Rao–Blackwellized approximation of this density (Robert & Casella, 2004) that furthermore converges at a parametric speed, as already noticed in Gelfand & Smith (1990):

$$\hat{p}(\theta^o|\mathbf{y}, G) = \frac{1}{M}\sum_{m=1}^{M} p(\theta^o|\mathbf{y}, \mathbf{z}^{(m)}, G),$$

where $\mathbf{z}^{(m)}, m = 1, \ldots, M$, are the posterior draws for the latent allocations $\mathbf{z} = (z_1, \ldots, z_n)$, introduced earlier in Chapter 1; see Chapter 5 for a review of posterior sampling methods.

However, for mixtures, the convergence of this estimate is very much hindered by the fact that it requires perfect symmetry in the Gibbs sampler, that is, complete label switching within the simulated Markov chain. When the completed chain $(z_1^{(m)}, \ldots, z_n^{(m)})$ remains instead concentrated around one single or a subset of the modes of the posterior distribution, the approximation of $\log \hat{p}(\theta^o|\mathbf{y}, G)$ based on Chib's representation fails, in that it is usually off by a numerical factor of order $\mathrm{O}(\log G!)$. Furthermore, this order cannot be used as a reliable correction, as noted by Neal (1999) and Frühwirth-Schnatter (2006).

A straightforward method of handling Markov chains that are not perfectly mixing (which is the usual setting) is found in Berkhof et al. (2003) (see also Frühwirth-Schnatter, 2006, Section 5.5.5; Lee et al., 2009) and can be interpreted as a form of Rao–Blackwellization. The proposed correction is to estimate $\hat{p}(\theta^o|\mathbf{y}, G)$ as an average computed over all possible permutations of the labels, thus forcing the label switching and the exchangeability of the labels to occur in a "perfect" manner. The new approximation can be expressed as

$$\tilde{p}(\theta^o|\mathbf{y}, G) = \frac{1}{MG!}\sum_{\mathfrak{s}\in\mathfrak{S}(G)}\sum_{m=1}^{M} p(\theta^o|\mathbf{y}, \mathfrak{s}(\mathbf{z}^{(m)}), G)\,,$$

[1] This was earlier called *the candidate's formula* by Julian Besag (1989).

where $\mathfrak{S}(G)$ traditionally denotes the set of the $G!$ permutations of $\{1,\ldots,G\}$ and where $\mathfrak{s}$ is one of those permutations. Note that the above correction can also be rewritten as

$$\tilde{p}(\theta^o|\mathbf{y},G) = \frac{1}{MG!}\sum_{\mathfrak{s}\in\mathfrak{S}(G)}\sum_{m=1}^{M} p(\mathfrak{s}(\theta^o)|\mathbf{y},\mathbf{z}^{(m)},G)\,, \tag{7.9}$$

as this may induce some computational savings. Further savings can be found in the importance sampling approach of Lee & Robert (2016), who reduce the number of permutations to be considered.

While Chib's representation has often been advocated as a reference method for computing the evidence, other methods abound within the literature, among them nested sampling (Skilling, 2007; Chopin & Robert, 2010), reversible jump MCMC (Green, 1995; Richardson & Green, 1997), particle filtering (Chopin, 2002), bridge sampling (Frühwirth-Schnatter, 2004) and path sampling (Gelman & Meng, 1998). Some of these methods are discussed next.

#### 7.2.3.2 Sampling-based approximations

If $G$ is moderate, sampling-based techniques are particularly useful for estimating the marginal likelihood of finite mixture models; see Frühwirth-Schnatter (2004) and Lee & Robert (2016). Frühwirth-Schnatter (2004) considered three such estimation techniques, namely importance sampling, reciprocal importance sampling, and bridge sampling.

For sampling-based techniques, one selects an importance density $q_G(\theta)$ which is easy to sample from and provides a rough approximation to the posterior density $p(\theta|\mathbf{y},G)$. Given a suitable importance density $q_G(\theta)$, an importance sampling approximation to the marginal likelihood is based on rewriting (7.2) as

$$p(\mathbf{y}|G) = \int \frac{p(\mathbf{y}|\theta,G)p(\theta|G)}{q_G(\theta)} q_G(\theta)d\theta.$$

Based on a sample $\theta^{(l)} \sim q_G(\theta)$, $l = 1,\ldots,L$, from the importance density $q_G(\theta)$, the importance sampling estimator of the marginal likelihood is given by

$$\hat{p}_{IS}(\mathbf{y}|G) = \frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y}|\theta^{(l)},G)p(\theta^{(l)}|G)}{q_G(\theta^{(l)})}. \tag{7.10}$$

Gelfand & Dey (1994) introduced reciprocal importance sampling, which is based on the observation that (7.8) can be written as

$$\frac{1}{p(\mathbf{y}|G)} = \frac{p(\theta|\mathbf{y},G)}{p(\mathbf{y}|\theta,G)p(\theta|G)}.$$

Integrating both sides of this equation with respect to the importance density $q_G(\theta)$ yields

$$\frac{1}{p(\mathbf{y}|G)} = \int \frac{q_G(\theta)}{p(\mathbf{y}|\theta,G)p(\theta|G)} p(\theta|\mathbf{y},G).$$

This leads to the reciprocal importance sampling estimator of the marginal likelihood, where the inverse of the ratio appearing in (7.10) is evaluated at the MCMC draws $\theta^{(m)}$, $m = 1,\ldots,M$, and no draws from the importance density $q_G(\theta)$ are required:

$$\hat{p}_{RI}(\mathbf{y}|G) = \left(\frac{1}{M}\sum_{m=1}^{M} \frac{q_G(\theta^{(m)})}{p(\mathbf{y}|\theta^{(m)},G)p(\theta^{(m)}|G)}\right)^{-1}.$$

These two estimators are special cases of bridge sampling (Meng & Wong, 1996):

$$p(\mathbf{y}|G) = \frac{\mathrm{E}_{q_G(\theta)}(\alpha(\theta)p(\mathbf{y}|\theta,G)p(\theta|G))}{\mathrm{E}_{p(\theta|\mathbf{y},G)}(\alpha(\theta)q_G(\theta))},$$

with specific functions $\alpha(\theta)$. The (formally) optimal choice for $\alpha(\theta)$ yields the bridge sampling estimator $\hat{p}_{BS}(\mathbf{y}|G)$ and combines draws $\theta^{(l)}$, $l = 1,\ldots,L$, from the importance density with MCMC draws $\theta^{(m)}$, $m = 1,\ldots,M$. Using $\hat{p}_{IS}(\mathbf{y}|G)$ as a starting value for $\hat{p}_{BS,0}(\mathbf{y}|G)$, the following recursion is applied until convergence to estimate $\hat{p}_{BS}(\mathbf{y}|G) = \lim_{t\to\infty} \hat{p}_{BS,t}(\mathbf{y}|G)$:

$$\hat{p}_{BS,t}(\mathbf{y}|G) = \frac{L^{-1}\sum_{l=1}^{L} \frac{p(\mathbf{y}|\theta^{(l)},G)p(\theta^{(l)}|G)}{Lq_G(\theta^{(l)}) + Mp(\mathbf{y}|\theta^{(l)},G)p(\theta^{(l)}|G)/\hat{p}_{BS,t-1}(\mathbf{y}|G)}}{M^{-1}\sum_{m=1}^{M} \frac{q_G(\theta^{(m)})}{Lq_G(\theta^{(m)}) + Mp(\mathbf{y}|\theta^{(m)},G)p(\theta^{(m)}|G)/\hat{p}_{BS,t-1}(\mathbf{y}|G)}}. \tag{7.11}$$

The reliability of these estimators depends on several factors. First, as shown by Frühwirth-Schnatter (2004), the tail behaviour of $q_G(\theta)$ compared to the mixture posterior $p(\theta|\mathbf{y},G)$ is relevant. Whereas the bridge sampling estimator $\hat{p}_{BS}(\mathbf{y}|G)$ is fairly robust to the tail behaviour of $q_G(\theta)$, $\hat{p}_{IS}(\mathbf{y}|G)$ is sensitive if $q_G(\theta)$ has lighter tails than $p(\theta|\mathbf{y},G)$, and $\hat{p}_{RI}(\mathbf{y}|G)$ is sensitive if $q_G(\theta)$ has fatter tails than $p(\theta|\mathbf{y},G)$. Second, as pointed out by Lee & Robert (2016), for any of these methods it is essential that the importance density $q_G(\theta)$ exhibits the same kind of multimodality as the mixture posterior $p(\theta|\mathbf{y},G)$ and all modes of the posterior density are covered by the importance density also for increasing values of $G$. Otherwise, sampling-based estimators of the marginal likelihood are prone to be biased for the same reason Chib's estimator is biased, as discussed in Section 7.2.3.1. A particularly stable estimator is obtained when bridge sampling is combined with a perfectly symmetric importance density $q_G(\theta)$. Before the various estimators are illustrated for three well-known data sets (Richardson & Green, 1997), we turn to the choice of appropriate importance densities.

*Importance densities for mixture analysis*

As manual tuning of the importance density $q_G(\theta)$ for each model under consideration is rather tedious, methods for choosing sensible importance densities in an unsupervised manner are needed. DiCiccio et al. (1997), for instance, suggested various methods to construct Gaussian importance densities from the MCMC output. However, the multimodality of the mixture posterior density with $G!$ equivalent modes evidently rules out such a simple choice. Frühwirth-Schnatter (1995) is one of the earliest references that used Rao–Blackwellization to construct an unsupervised importance density from the MCMC output to compute marginal likelihoods via sampling-based approaches and applied this idea to model selection for linear Gaussian state space models. Frühwirth-Schnatter (2004) extends this idea to finite mixture and Markov switching models where the complete-data posterior $p(\theta|\mathbf{y},\mathbf{z})$ is available in closed form. Lee & Robert (2016) discuss importance sampling schemes based on (nearly) perfectly symmetric importance densities.

For a mixture distribution, where the component-specific parameters $\theta_g$ can be sampled in one block from the complete-data posterior $p(\theta_g|\mathbf{z},\mathbf{y})$, Rao–Blackwellization yields the importance density

$$q_G(\theta) = \frac{1}{S}\sum_{s=1}^{S} p(\eta|\mathbf{z}^{(s)}) \prod_{g=1}^{G} p(\theta_g|\mathbf{z}^{(s)},\mathbf{y}), \tag{7.12}$$

where $\mathbf{z}^{(s)}$ are the posterior draws for the latent allocations. The construction of this importance density is fully automatic and it is sufficient to store the moments of these conditional densities (rather than the allocations $\mathbf{z}$ themselves) during MCMC sampling for later evaluation. This method can be extended to cases where sampling $\theta_g$ from $p(\theta_g|\mathbf{z},\mathbf{y})$ requires two (or even more) blocks such as for Gaussian mixtures where $\theta_g = (\mu_g, \sigma_g^2)$ is sampled in two steps from $p(\mu_g|\sigma_g^2,\mathbf{z},\mathbf{y})$ and $p(\sigma_g^2|\mu_g,\mathbf{z},\mathbf{y})$.

Concerning the number of components in (7.12), on the one hand $S$ should be small for computational reasons, because $q_G(\theta)$ has to be evaluated for each of the $S$ components numerous times (e.g. $L$ times for the importance sampling estimator (7.10)). On the other hand, as mentioned above, it is essential that $q_G(\theta)$ covers all symmetric modes of the mixture posterior, and this will require a dramatically increasing number of components $S$ as $G$ increases. Hence, any of these estimators is limited to moderate values of $G$, say up to $G = 6$.

Various strategies are available to ensure multimodality in the construction of the importance density. Frühwirth-Schnatter (2004) chooses $S = M$ and relies on random permutation Gibbs sampling (Frühwirth-Schnatter, 2001) by applying a randomly selected permutation $\mathfrak{s}_m \in \mathfrak{S}(G)$ at the end of the $m$th MCMC sweep to define a permutation $\mathbf{z}^{(s)} = \mathfrak{s}_m(\mathbf{z}^{(m)})$ of the posterior draw $\mathbf{z}^{(m)}$ of the allocation vector. The random permutations $\mathfrak{s}_1, \ldots, \mathfrak{s}_M$ guarantee multimodality of $q_G(\theta)$ in (7.12); however, as discussed above, it is important to ensure good mixing of the underlying permutation sampler over all $G!$ equivalent posterior modes. Only if $S$ is large compared to $G!$ are all symmetric modes visited by random permutation sampling. Choosing, for instance, $S = S_0 G!$ ensures that each mode is visited on average $S_0$ times.

As an alternative to random permutation sampling, approaches exploiting full permutations have been suggested; see, for example, Frühwirth-Schnatter (2004). Importance sampling schemes exploiting full permutation were discussed in full detail in Lee & Robert (2016). The definition of a fully symmetric importance density $q_G(\theta)$ is related to the correction for Chib's estimator discussed earlier in (7.9):

$$q_G(\theta) = \frac{1}{S_0 G!} \sum_{\mathfrak{s} \in \mathfrak{S}(G)} \sum_{s=1}^{S_0} p(\eta|\mathfrak{s}(\mathbf{z}^{(s)})) \prod_{g=1}^{G} p(\theta_g|\mathfrak{s}(\mathbf{z}^{(s)}),\mathbf{y}). \tag{7.13}$$

This construction, which has $S = S_0 G!$ components, is based on a small number $S_0$ of particles $\mathbf{z}^{(s)}$, as $q_G(\theta)$ needs to be only a rough approximation to the mixture posterior $p(\theta|\mathbf{y}, G)$ and estimators such as bridge sampling will be robust to the tail behaviour of $q_G(\theta)$. In (7.13), all symmetric modes are visited exactly $S_0$ times. The moments of the $S_0$ conditional densities need to be stored for only one of the $G!$ permutations and, again, this construction can be extended to the case where the components of $\theta_g$ are sampled in more than one block. Lee & Robert (2016) discuss strategies for reducing the computational burden associated with evaluating $q_G(\theta)$.

Frühwirth-Schnatter (2006, p. 146) and Lee & Robert (2016) discuss a simplified version of (7.13) where the random sequence $\mathbf{z}^{(s)}$, $s = 1, \ldots, S_0$, is substituted by a single optimal partition $\mathbf{z}^\star$ such as the maximum *a posteriori* (MAP) estimator:

$$q_G(\theta) = \frac{1}{G!} \sum_{\mathfrak{s} \in \mathfrak{S}(G)} p(\theta|\mathfrak{s}(\mathbf{z}^\star),\mathbf{y}).$$

In MATLAB, the bayesf package (Frühwirth-Schnatter, 2018) allows one to estimate $\hat{p}_{BS}(\mathbf{y}|G)$, $\hat{p}_{IS}(\mathbf{y}|G)$ and $\hat{p}_{RI}(\mathbf{y}|G)$ with the importance density being constructed either as in (7.12) using random permutation sampling or as in (7.13) using full permutation sampling.

*Example: Marginal likelihoods for the data sets in Richardson & Green (1997)*

By way of illustration, marginal likelihoods are computed for mixtures of $G$ univariate normal distributions $\mathcal{N}(\mu_g, \sigma_g^2)$ for $G = 2, \ldots, 6$ for the acidity data, the enzyme data and the galaxy data studied by Richardson & Green (1997) in the framework of reversible jump MCMC (see Section 7.4.2 for a short description of this one-sweep method). We use the same priors as Richardson & Green, namely the symmetric Dirichlet prior $\eta \sim \mathcal{D}_G(1)$, the normal prior $\mu_g \sim \mathcal{N}(m, R^2)$, the inverse gamma prior $\sigma_g^2 \sim \mathcal{IG}(2, C_0)$ and the gamma prior $C_0 \sim \mathcal{G}(0.2, 10/R^2)$, where $m$ and $R$ are the midpoint and the length of the observation interval. For a given $G$, full conditional Gibbs sampling is performed for $M = 12{,}000$ draws after a burn-in of 2000, by iteratively sampling from $p(\sigma_g^2|\mu_g, C_0, \mathbf{z}, \mathbf{y})$, $p(\mu_g|\sigma_g^2, \mathbf{z}, \mathbf{y})$, $p(C_0|\sigma_1^2, \ldots, \sigma_G^2)$, $p(\eta|\mathbf{z})$ and $p(\mathbf{z}|\theta, \mathbf{y})$.

A fully symmetric importance density $q_{G,F}(\theta)$ is constructed from (7.13), where $S_0 = 100$ components are selected for each mode. For comparison, an importance density $q_{G,R}(\theta)$ is constructed from (7.12) with $S = S_0 G!$, ensuring that for random permutation sampling each mode is visited on average $S_0$ times. However, unlike $q_{G,F}(\theta)$, the importance density $q_{G,R}(\theta)$ is not fully symmetric. Ignoring the dependence between $\mu_g$ and $\sigma_g^2$, the component densities are constructed from conditionally independent densities, given the $s$th draw of $(\mathbf{z}, \theta_1, \ldots, \theta_G, C_0)$:

$$p(\mu_g, \sigma_g^2|\mathbf{z}^{(s)}, \theta_g^{(s)}, C_0^{(s)}, \mathbf{y}) = p(\mu_g|\sigma_g^{2,(s)}, \mathbf{z}^{(s)}, \mathbf{y}) p(\sigma_g^2|\mu_g^{(s)}, C_0^{(s)}, \mathbf{z}^{(s)}, \mathbf{y}).$$

Prior evaluation is based on the marginal prior $p(\sigma_1^2, \ldots, \sigma_G^2)$, where $C_0$ is integrated out.

This yields in total six estimators, $\hat{p}_{BS,F}(\mathbf{y}|G)$, $\hat{p}_{IS,F}(\mathbf{y}|G)$ and $\hat{p}_{RI,F}(\mathbf{y}|G)$ for full permutation sampling and $\hat{p}_{BS,R}(\mathbf{y}|G)$, $\hat{p}_{IS,R}(\mathbf{y}|G)$ and $\hat{p}_{RI,R}(\mathbf{y}|G)$ for random permutation sampling, for each $G = 2, \ldots, 6$. For each estimator, standard errors SE are computed as in Frühwirth-Schnatter (2004). Results are visualized in Figure 7.3, by plotting all six estimators $\hat{p}_\bullet(\mathbf{y}|G)$ as well as $\hat{p}_\bullet(\mathbf{y}|G) \pm 3\,\mathrm{SE}$ over $G$ for all three data sets. Good estimators should be unbiased with small standard errors and the order in which the six estimators are arranged (which is the same for all $G$s) is related to this quality measure.

There is a striking difference in the reliability of the six estimators, in particular as $G$ increases. Reciprocal importance sampling is particularly unreliable and the estimated values of $\log \hat{p}_{RI,R}(\mathbf{y}|G)$ under $q_{G,R}(\theta)$ tend to be extremely biased for $G \geq 4$, even if the bias is reduced to a certain extent by choosing the fully symmetric importance density $q_{G,F}(\theta)$. Also the two other estimators $\log \hat{p}_{IS,R}(\mathbf{y}|G)$ and $\log \hat{p}_{BS,R}(\mathbf{y}|G)$ tend to be biased under $q_{G,R}(\theta)$, and bridge sampling is more sensitive than importance sampling to choosing an importance density that is not fully symmetric.

Unlike for reciprocal importance sampling, the bias disappears for both bridge sampling and importance sampling under the fully symmetric importance density $q_{G,F}(\theta)$, and $\log \hat{p}_{IS,F}(\mathbf{y}|G)$ and $\log \hat{p}_{BS,F}(\mathbf{y}|G)$ yield more or less identical results. However, due to the robustness of bridge sampling with respect to the tail behaviour of $q_{G,F}(\theta)$, we find that the standard errors of $\log \hat{p}_{BS,F}(\mathbf{y}|G)$ are often considerably smaller than the standard errors of $\log \hat{p}_{IS,F}(\mathbf{y}|G)$, in particular for the enzyme data.

Based on $\log \hat{p}_{BS,F}(\mathbf{y}|G)$, marginal likelihood evaluation yields the following results for the three data sets. For the acidity data, $\log \hat{p}_{BS,F}(\mathbf{y}|G = 3) = -198.2$ and $\log \hat{p}_{BS,F}(\mathbf{y}|G = 4) = -198.3$ are more less the same, with the log odds of $G = 3$ over $G = 4$ being equal to 0.1. Also for the enzyme data, with $\log \hat{p}_{BS,F}(\mathbf{y}|G = 3) = -74.2$ and $\log \hat{p}_{BS,F}(\mathbf{y}|G = 4) = -74.3$, the log odds of $G = 3$ over $G = 4$ are equal to 0.1. Finally, for the galaxy data, $\log \hat{p}_{BS,F}(\mathbf{y}|G = 5) = \log \hat{p}_{BS,F}(\mathbf{y}|G = 6) = -225.9$. Hence, under the prior $p(\theta|G)$ employed by Richardson & Green (1997), for all three data sets no clear distinction can be made between two values of $G$ based on the marginal likelihood. However, if the marginal likelihoods are combined with a prior on the number of components such as $G - 1 \sim \mathcal{P}(1)$

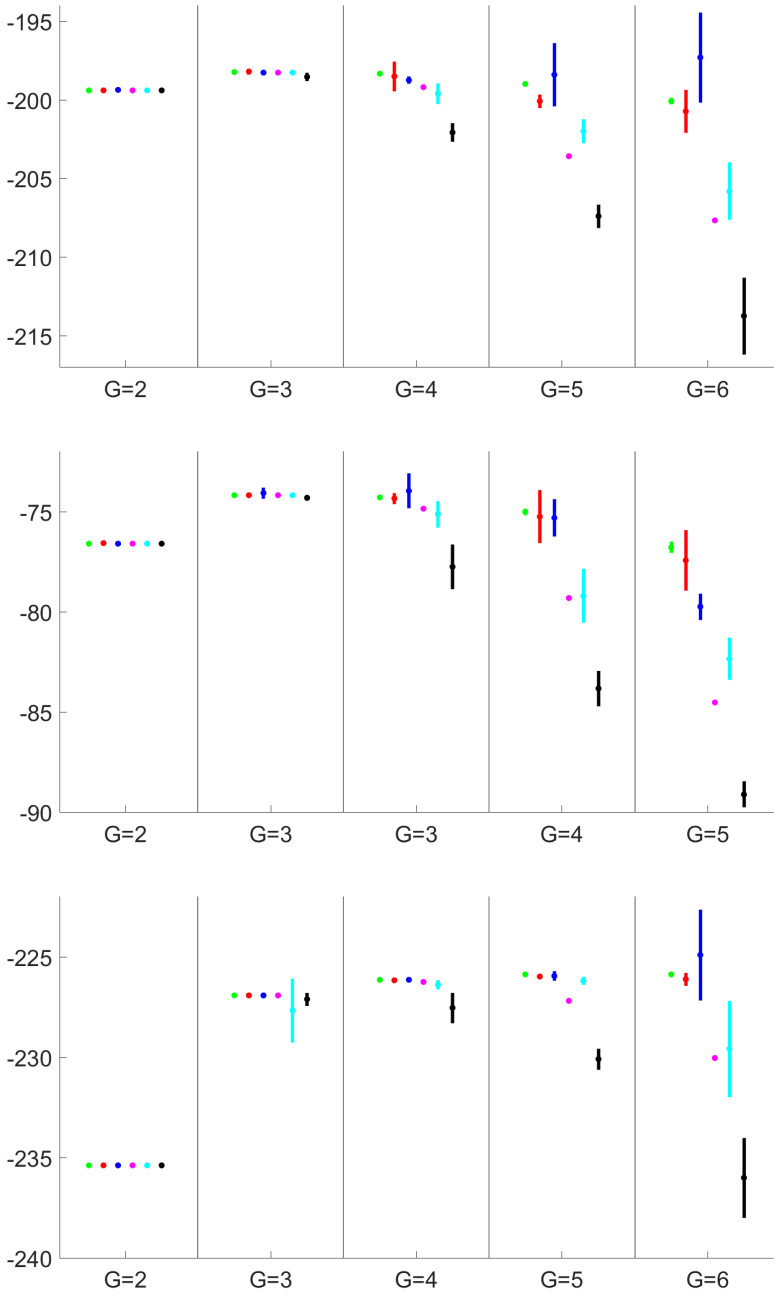


**FIGURE 7.3**
Marginal likelihood estimation for the benchmarks in Richardson & Green (1997): the acidity data (top), the enzyme data (middle) and the galaxy data (bottom) over $G = 2, \ldots, G = 6$. For each $G$, six estimators $\hat{p}_\bullet(\mathbf{y}|G)$ are given together with $\hat{p}_\bullet(\mathbf{y}|G) \pm 3\,\mathrm{SE}$ in the order $\hat{p}_{BS,F}(\mathbf{y}|G)$, $\hat{p}_{IS,F}(\mathbf{y}|G)$, $\hat{p}_{IS,R}(\mathbf{y}|G)$, $\hat{p}_{BS,R}(\mathbf{y}|G)$, $\hat{p}_{RI,F}(\mathbf{y}|G)$ and $\hat{p}_{RI,R}(\mathbf{y}|G)$ from left to right.

(Nobile, 2004), then the log posterior odds, being equal to 1.5 for the acidity and the enzyme data and 1.8 for the galaxy data, yield evidence for the smaller of the two values of $G$ for all three data sets.

## 7.3 Selecting $G$ in the Framework of Model-Based Clustering

Assuming that the data stem from one of the models under comparison is most often unrealistic and can be misleading when using the AIC or BIC. Now a common feature of standard penalized likelihood criteria is that they abstain from taking the modelling purpose into account, except when inference is about estimating the data density. In particular, misspecification can lead to overestimating the complexity of a model in practical situations. Taking the modelling purpose into account when selecting a model leads to alternative model selection criteria that favor useful and parsimonious models. This viewpoint is particularly relevant when considering a mixture model for model-based clustering; see Chapter 8 for a review of this important application of mixture models.

### 7.3.1 Mixtures as partition models

Clustering arises in a natural way when an i.i.d. sample is drawn from the finite mixture distribution (7.1) with weights $\eta = (\eta_1, \ldots, \eta_G)$. As explained in Chapter 1, Section 1.1.3, each observation $y_i$ can be associated with the component, indexed by $z_i$, that generated this data point:

$$\begin{aligned} z_i|\eta &\sim \mathcal{M}(1, \eta_1, \ldots, \eta_G), \\ y_i|z_i &\sim f_{z_i}(y_i|\theta_{z_i}). \end{aligned} \tag{7.14}$$

Let $\mathbf{z} = (z_1, \ldots, z_n)$ be the collection of all component indicators that were used to generate the $n$ data points $\mathbf{y} = (y_1, \ldots, y_n)$. Obviously, $\mathbf{z}$ defines a partition of the data. A cluster $C_g = \{i|z_i = g\}$ is thus defined as a subset of the data indices $\{1, \ldots, n\}$, containing all observations with identical allocation variables $z_i$. Hence, the indicators $\mathbf{z}$ define a partition $\mathcal{C} = \{C_1, \ldots, C_{G_+}\}$ of the $n$ data points, where $y_i$ and $y_j$ belong to the same cluster if and only if $z_i = z_j$. The partition $\mathcal{C}$ contains $G_+ = |\mathcal{C}|$ clusters, where $|\mathcal{C}|$ is the cardinality of $\mathcal{C}$. In a Bayesian context, finite mixture models imply *random partitions* over the lattice

$$\mathcal{S}_G^n = \{(z_1, \ldots, z_n) : z_i \in \{1, \ldots, G\}, i = 1, \ldots, n\},$$

as will be discussed in detail in Section 7.3.3.

In model-based clustering, a finite mixture model is applied to recover the (latent) allocation indicators $\mathbf{z}$ from the data and to estimate a suitable partition of the data. A useful quantity in this respect is the so-called fuzzy classification matrix $\tau$. The elements $\tau_{ig}$, with $i = 1, \ldots, n$ and $g = 1, \ldots, G$, of $\tau$ are equal to the conditional probability that observation $y_i$ arises from component $g$ in a mixture model of order $G$ given $y_i$:

$$\tau_{ig} = \mathrm{P}(z_i = g|y_i, \theta) = \mathrm{P}(z_{ig} = 1|y_i, \theta) = \frac{\eta_g f_g(y_i \mid \theta_g)}{\sum_{j=1}^{G} \eta_j f_j(y_i \mid \theta_j)}, \tag{7.15}$$

where $z_{ig} = \mathbb{I}(z_i = g)$. The entropy $\mathrm{ENT}(\theta; G)$ corresponding to a fuzzy classification matrix $\tau$ is defined as

$$\mathrm{ENT}(\theta; G) = -\sum_{g=1}^{G} \sum_{i=1}^{n} \tau_{ig} \log \tau_{ig} \geq 0. \tag{7.16}$$

Both $\tau$ and $\mathrm{ENT}(\theta; G)$ are data-driven measures of the ability of a $G$-component mixture model to provide a relevant partition of the data. If the mixture components are well separated for a given $\theta$, then the classification matrix $\tau$ tends to define a clear partition of the data set $\mathbf{y} = (y_1, \ldots, y_n)$, with $\tau_{ig}$ being close to 1 for one component and close to 0 for all other components. In this case, $\mathrm{ENT}(\theta; G)$ is close to 0. On the other hand, if the mixture components are poorly separated, then $\mathrm{ENT}(\theta; G)$ takes values larger than zero. The maximum value $\mathrm{ENT}(\theta; G)$ can take is $n \log G$, which is the entropy of the uniform distribution which assigns $y_i$ to all $G$ clusters with the same probability $\tau_{ig} \equiv 1/G$.

In a Bayesian context, the fuzzy classification matrix is instrumental for joint estimation of the parameter $\theta$ and $\mathbf{z}$ within Gibbs sampling using data augmentation (see, for example, Robert & Casella, 2004). In a frequentist framework, the estimated classification matrix $\hat{\tau}$, given a suitable estimate $\hat{\theta}_G$ of the mixture parameters $\theta$ (e.g. the MLE), can be used to derive an estimator $\hat{\mathbf{z}}$ of the partition of the data; see also Chapter 8, Section 8.2.4. As will be discussed in Section 7.3.2, the entropy of the estimated classification matrix $\hat{\tau}$ plays an important role in defining information criteria for choosing $G$ in a clustering context.

### 7.3.2 Classification-based information criteria

As discussed in Section 7.2.2.1 within the framework of density estimation, the BIC enjoys several desirable properties; however, within cluster analysis it shows a tendency to overestimate $G$; see, for instance, Celeux & Soromenho (1996). The BIC does not take the clustering purposes for assessing $G$ into account, regardless of the separation of the clusters. To overcome this limitation, an attractive possibility is to select $G$ so that the resulting mixture model leads to the clustering of the data with the largest evidence. This is the purpose of various classification-based information criteria such as the integrated complete-data likelihood criterion that are discussed in this subsection.

In a classification context, it is useful to state a simple relation linking the log of the observed-data density $p(\mathbf{y}|\theta)$ and the complete-data density $p(\mathbf{y}, \mathbf{z}|\theta)$. The observed-data log likelihood of $\theta$ for a sample $\mathbf{y}$, denoted by $\log \ell_o(\theta; G)$, is given by

$$\log \ell_o(\theta; G) = \sum_{i=1}^{n} \log \left[ \sum_{g=1}^{G} \eta_g f_g(y_i \mid \theta_g) \right],$$

whereas the complete-data log likelihood of $\theta$ for the complete sample $(\mathbf{y}, \mathbf{z})$, denoted by $\log \ell c(\theta; G)$, reads

$$\log \ell c(\theta, \mathbf{z}; G) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(\eta_g f_g(y_i \mid \theta_g)),$$

where $z_{ig} = \mathbb{I}(z_i = g)$, $g = 1, \ldots, G$. These log likelihoods are linked in the following way:

$$\log \ell c(\theta, \mathbf{z}; G) = \log \ell_o(\theta; G) - \mathrm{EC}(\theta, \mathbf{z}; G), \tag{7.17}$$

where

$$\mathrm{EC}(\theta, \mathbf{z}; G) = -\sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log \tau_{ig} \geq 0.$$

Since $\mathrm{E}(z_{ig}|\theta, y_i) = \mathrm{P}(z_{ig} = 1|\theta, y_i) = \tau_{ig}$, we obtain that the expectation of $\mathrm{EC}(\theta, \mathbf{z}; G)$ with respect to the conditional distribution $p(\mathbf{z}|\mathbf{y}, \theta)$ for a given $\theta$ is equal the entropy $\mathrm{ENT}(\theta; G)$ defined in (7.16). Hence, the entropy can be regarded as a penalty for the observed-data likelihood in cases where the resulting clusters are not well separated.

#### 7.3.2.1 The integrated complete-data likelihood criterion

The integrated (complete-data) likelihood related to the complete data $(\mathbf{y}, \mathbf{z})$ is

$$p(\mathbf{y}, \mathbf{z} \mid G) = \int_{\Theta_G} p(\mathbf{y}, \mathbf{z} \mid G, \theta) p(\theta \mid G) d\theta,$$

where

$$p(\mathbf{y}, \mathbf{z} \mid G, \theta) = \prod_{i=1}^{n} p(y_i, z_i \mid G, \theta) = \prod_{i=1}^{n} \prod_{g=1}^{G} \eta_g^{z_{ig}} \left[ f_g(y_i \mid \theta_g) \right]^{z_{ig}} .$$

This integrated complete-data likelihood (ICL) takes the missing data $\mathbf{z}$ into account and can be expected to be relevant for choosing $G$ in a clustering context. However, computing the ICL is challenging for various reasons. First, computing the ICL involves an integration in high dimensions. Second, the labels $\mathbf{z}$ are unobserved (missing) data. To approximate the ICL, a BIC-like approximation is possible (Biernacki et al., 2000):

$$\log p(\mathbf{y}, \mathbf{z} \mid G) \approx \log p(\mathbf{y}, \mathbf{z} \mid G, \hat{\theta}^{\mathbf{z}}) - \frac{\upsilon_G}{2} \log n,$$

where

$$\hat{\theta}^{\mathbf{z}} = \arg\max_{\theta} p(\mathbf{y}, \mathbf{z} \mid G, \theta),$$

and $\upsilon_G$ is the number of free parameters of the mixture model $\mathcal{M}_G$. Note that this approximation involves the complete-data likelihood, $\mathrm{L_c}(\theta, \mathbf{z}; G) = p(\mathbf{y}, \mathbf{z} \mid G, \theta)$; however, $\mathbf{z}$ and, consequently, $\hat{\theta}^{\mathbf{z}}$ are unknown. First, approximating $\hat{\theta}^{\mathbf{z}} \approx \hat{\theta}_G$, with $\hat{\theta}_G$ being the MLE of the $G$-component mixture parameter $\theta$, is expected to be valid for well-separated components. Second, given $\hat{\theta}_G$, the missing data $\mathbf{z}$ are imputed using the MAP estimator $\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\theta}_G)$ defined by

$$\hat{z}_{ig} = \begin{cases} 1, & \text{if } \operatorname{argmax}_l \tau_{il}(\hat{\theta}_G) = g, \\ 0, & \text{otherwise.} \end{cases}$$

This leads to the criterion

$$\mathrm{ICLbic}(G) = \log p(\mathbf{y}, \hat{\mathbf{z}} \mid G, \hat{\theta}_G) - \frac{\upsilon_G}{2} \log n.$$

Exploiting (7.17), one obtains that the ICLbic criterion takes the form of a BIC criterion, penalized by the estimated entropy

$$\mathrm{ENT}(\hat{\theta}_G; G) = -\sum_{g=1}^{G} \sum_{i=1}^{n} \hat{\tau}_{ig} \log \hat{\tau}_{ig} \geq 0,$$

with $\hat{\tau}_{ig}$ denoting the conditional probability that $y_i$ arises from the $g$th mixture component ($i = 1, \ldots, n$, $g = 1, \ldots, G$) under the parameter $\hat{\theta}_G$; see (7.15).

Because of this additional entropy term, the ICLbic criterion favours values of $G$ giving rise to partitions of the data with the highest evidence. In practice, the ICLbic appears to provide a stable and reliable estimation of $G$ for real data sets and also for simulated data sets from mixtures when the components do not overlap too much. However, it should be noted that the ICLbic, which is not concerned with discovering the true number of mixture components, can underestimate the number of components for simulated data arising from mixtures with poorly separated components.

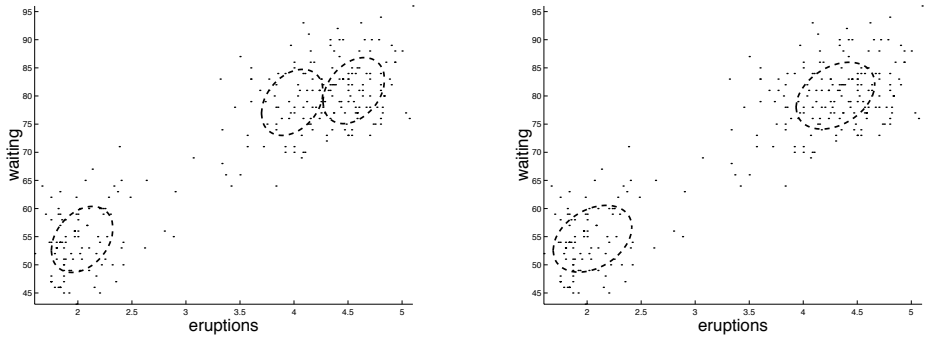


**FIGURE 7.4**
Cluster ellipses for the Old Faithful Geyser data: (left) the BIC solution; (right) the ICLbic solution.

*An illustrative comparison of the BIC and ICLbic*

Obviously, in many situations where the mixture components are well separated, the BIC and ICLbic select the same number of mixture components. But the following small numerical example aims to illustrate a situation where these two criteria give different answers.

We start from a benchmark (genuine) data set known as the *Old Faithful Geyser*. Each of the 272 observations consists of two measurements: the duration of the eruption and the waiting time before the next eruption of the Old Faithful Geyser, in Yellowstone National Park, USA. We consider a bivariate Gaussian mixture model with component densities $\mathcal{N}(\mu_k, \Sigma_k)$ with unconstrained covariance matrices $\Sigma_k$.

For this data set, Figure 7.4 shows that the ICLbic selects with a large evidence $G = 2$, while the BIC slightly prefers $G = 3$ to $G = 2$. The BIC solution with $G = 3$ components appears to model deviations from normality in one of the two obvious clusters, rather than a relevant additional cluster.

#### 7.3.2.2 The conditional classification likelihood

In a model-based clustering context where a cluster is associated with a mixture component, it is sensible in view of (7.17) to maximize the conditional expectation of the complete-data log likelihood (Baudry, 2015),

$$\log \mathrm{L}_{\mathrm{cc}}(\theta; G) = \mathrm{E}_{\mathbf{z}}(\log \ell c(\theta, \mathbf{z}; G)) = \log \ell_o(\theta; G) - \mathrm{ENT}(\theta; G),$$

rather than the observed-data log likelihood function $\log \ell_o(\theta; G)$. This can be done through an EM-type algorithm where the M step at iteration $s + 1$ involves finding

$$\theta^{(s+1)} \in \underset{\theta \in \Theta_G}{\operatorname{argmax}} \left( \log \ell_o(\theta; G) + \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(s)} \log \tau_{ig} \right), \tag{7.18}$$

where the $\tau_{ig}$ are defined as in (7.15) and

$$\tau_{ig}^{(s)} = \frac{\eta_g^{(s)} f_g(y_i \mid \theta_g^{(s)})}{\sum_{j=1}^{G} \eta_j^{(s)} f_j(y_i \mid \theta_j^{(s)})}.$$

This M step can be performed by using an adaptation of the so-called Bayesian expectation maximization (BEM) of Lange (1999). The resulting algorithm inherits the fundamental

property of EM to increase the criterion $\log \mathrm{L_{cc}}(\theta)$, which does not depend on $\mathbf{z}$, at each iteration.

In this context, Baudry (2015) considered choosing $G$ from a penalized criterion of the form

$$\mathrm{L_{cc}\text{-}ICL}(G) = -\log \mathrm{L_{cc}}(\widehat{\theta}_G^{\mathrm{MLccE}}; G) + \mathrm{pen}(G),$$

where $\widehat{\theta}_G^{\mathrm{MLccE}} = \arg\max_\theta \log \mathrm{L_{cc}}(\theta; G)$. Under regularity conditions and the standard condition for information criteria the following holds. Assuming that $\mathrm{pen} : \{1, \ldots, G_{\max}\} \to \mathbb{R}^+$ satisfies

$$\begin{cases} \mathrm{pen}(G) = o_{\mathbb{P}}(n), \quad \text{as } n \to \infty, \\ \big(\mathrm{pen}(G) - \mathrm{pen}(G')\big) \xrightarrow[n\to\infty]{\mathbb{P}} \infty, \quad \text{if } G' < G, \end{cases}$$

and defining $\hat{G} = \min \operatorname{argmin}_G \mathrm{L_{cc}\text{-}ICL}(G)$, Baudry (2015) proved that

$$\mathbb{P}\big[\widehat{G} \neq G_0\big] \xrightarrow[n\to\infty]{} 0,$$

where $G_0$ is the minimum number of components such that the bias of the models is stationary for $G \geq G_0$,

$$G_0 = \min \operatorname*{argmax}_G \mathrm{E}_{p_0}\big[\log \ell c(\theta_G^0)\big],$$

with

$$\theta_G^0 = \operatorname*{argmin}_{\theta \in \Theta_G} \Big\{ d_{\mathrm{KL}}\big(p_0, p(\,.\,;\theta)\big) + E_{p_0}\big[\mathrm{EC}(\theta; G)\big] \Big\},$$

$d_{\mathrm{KL}}\big(p_0, p(\,.\,;\theta)\big)$ being the Kullback–Leibler distance between the true distribution $p_0$ of the data and the mixture distribution with parameter $\theta$. Moreover, Baudry (2015) deduces that, by analogy with the BIC, an interesting identification criterion to be minimized is

$$\mathrm{L_{cc}\text{-}ICL}(G) = -\log \mathrm{L_{cc}}(\widehat{\theta}_G^{\mathrm{MLccE}}; G) + \frac{\upsilon_G}{2} \log n.$$

The criterion ICLbic can thus be viewed as an approximation of $\mathrm{L_{cc}}$-ICL. Therefore, the criterion $\mathrm{L_{cc}}$-ICL underlies a notion of class that is a compromise between the "mixture component" and the "cluster" points of view.

#### 7.3.2.3 Exact derivation of the ICL

Like the BIC, the ICL has been defined in a Bayesian framework, but its asymptotic approximations ICLbic and $\mathrm{L_{cc}}$-ICL are not intrinsically Bayesian, since they do not depend on the associated prior distribution. However, if the mixture components belong to the exponential family, it is possible to get closed-form expressions for the ICL (see Biernacki et al., 2010, or Bertoletti et al., 2015). With such closed-form expressions, it is possible to compute the ICL values by replacing the missing labels $\mathbf{z}$ with their most probable values using the MAP operator after estimating the parameter $\hat{\theta}_G$ as the posterior mode or the MLE (see Biernacki et al., 2010). An alternative is to optimize the exact ICL in $\mathbf{z}$. The limitations of approaches based on exact ICL computing are twofold.

*Choosing non-informative prior distributions*

Except for categorical data which involve mixtures of multivariate discrete distributions, there is no proper consensual non-informative prior distribution for other classes of mixture models such as Gaussian or Poisson mixture models (see Chapter 4). It is obviously possible to choose exchangeable weakly informative hyperparameters with conjugate prior distributions for the parameters of the mixture components. However, the posterior distribution and thus the resulting ICL values will inevitably depend on these hyperparameters.

For the latent class model on categorical data, deriving the exact ICL is easier, since the non-informative conjugate Dirichlet prior distributions $\mathcal{D}_G(e_0)$ are proper for the weight distribution of the mixture. Following the recommendation of Frühwirth-Schnatter (2011), it has been demonstrated that choosing $e_0 = 4$ is expected to provide a stable selection of $G$ (see, for instance, Keribin et al., 2015). Numerical experiments on simulated data proved that exact ICL computed with plug-in estimates $\hat{\theta}_G$ of the parameter could provide different and more reliable estimation of $G$ than the ICLbic for small sample sizes. Thus, when conjugate non-informative prior distributions are available, deriving a non-asymptotic approximation of ICL can be feasible.

*Optimizing the exact ICL*

Several authors have considered the direct optimization of the exact ICL in $\mathbf{z}$ without estimating $\theta$. Bertoletti et al. (2015), Côme & Latouche (2015) and Wyse et al. (2017) have proposed greedy algorithms, while Tessier et al. (2006) proposed using evolutionary optimization algorithms. At this point, it is important to remark that the optimization problem has to be solved in a search space with about $O(G_{\max}^n)$ elements, where $G_{\max}$ is the maximum number of components allowed. This means that the optimization problem becomes quite formidable for $n$ large. In addition, the proposed greedy algorithms are highly sensitive to the numerous local optima and have only been experimented with for moderate sample sizes. This is the reason why evolutionary algorithms are expected to be useful but they need to be calibrated (to choose the tuning parameters) and are expensive in computing time.

### 7.3.3 Bayesian clustering

In the context of Bayesian clustering (see Lau & Green, 2007, for an excellent review), where the allocation indicator $\mathbf{z} = (z_1, \ldots, z_n)$ is regarded as a latent variable, a finite mixture model implies *random partitions* over the lattice $\mathcal{S}_G^n$. Hence, for a given order $G$ of the mixture distribution (7.1), both the prior density $p(\mathbf{z}|G)$ and the posterior density $p(\mathbf{z}|G, \mathbf{y})$ are discrete distributions over the lattice $\mathcal{S}_G^n$. Although this induces a change of prior modelling, Lau & Green (2007) discuss Bayesian nonparametric (BNP; see Chapter 6) methods to estimate the number of clusters. We discuss the BNP perspective further in Section 7.4.4 and refer to Chapter 6 for a comprehensive treatment.

For a finite mixture model, the Dirichlet prior $\eta \sim \mathcal{D}(e_1, \ldots, e_G)$ on the weight distribution strongly determines what the prior distribution $p(\mathbf{z}|G)$ looks like. To preserve symmetry with respect to relabelling, typically the symmetric Dirichlet prior $\mathcal{D}_G(e_0)$ is employed, where $e_1 = \ldots = e_G = e_0$. The corresponding prior $p(\mathbf{z}|G) = \int \prod_{i=1}^n p(z_i|\eta) d\,\eta$ is given by

$$p(\mathbf{z}|G) = \frac{\Gamma(Ge_0)}{\Gamma(n + Ge_0)\Gamma(e_0)^{G_+}} \prod_{g:n_g>0} \Gamma(n_g + e_0), \tag{7.19}$$

where $n_g = \sum_{i=1}^n \mathbb{I}(z_i = g)$ is the number of observations in cluster $g$ and $G_+$ is defined as the number of non-empty clusters,

$$G_+ = G - \sum_{g=1}^{G} \mathbb{I}(n_g = 0). \tag{7.20}$$

As mentioned earlier, in model-based clustering interest lies in estimating the number of clusters $G_+$ in the $n$ data points rather than the number of components $G$ of the mixture

distribution (7.1), and it is important to distinguish between both quantities. Only a few papers make this clear distinction between the number of mixture components $G$ and the number of data cluster $G_+$ for finite mixture models (Nobile, 2004; Miller & Harrison, 2018; Malsiner-Walli et al., 2017; Frühwirth-Schnatter & Malsiner-Walli, 2018).

A common criticism concerning the application of finite mixture models in a clustering context is that the number of components $G$ needs to be known *a priori*. However, what is yet not commonly understood is (a) that the really relevant question is whether or not the *number of clusters* $G_+$ in the data is known *a priori* and (b) that even a finite mixture with a fixed value of $G$ can imply a random prior distribution on $G_+$. By way of further illustration, let $n_g = \sum_{i=1}^{n} \mathbb{I}(z_i = g)$ be the number of observations generated by the components $g = 1, \ldots, G$. Then (7.14) implies that $n_1, \ldots, n_G$ follow a multinomial distribution:

$$n_1, \ldots, n_G | \eta \sim \mathcal{M}(n, \eta_1, \ldots, \eta_G). \tag{7.21}$$

Depending on the weights $\eta = (\eta_1, \ldots, \eta_G)$ appearing in the mixture distribution (7.1), multinomial sampling according to (7.21) may lead to partitions with $n_g$ being zero, leading to so-called "empty components". In this case, fewer than $G$ mixture components were used to generate the $n$ data points which contain $G_+$ non-empty clusters, where $G_+$ is defined as in (7.20).

In a Bayesian framework towards finite mixture modelling, the Dirichlet prior $\eta \sim \mathcal{D}_G(e_0)$ on the component weights controls whether, *a priori*, $G_+$ is equal to $G$ and no empty components occur. In particular, if $e_0$ is close to 0, then $G_+$ is a random variable taking *a priori* values smaller than $G$ with high probability. Exploiting the difference between $G_+$ and $G$ in an overfitting mixture with a prior on the weight distribution that strongly shrinks redundant component weights towards 0 is a cornerstone of the concept of sparse finite mixtures (Malsiner-Walli et al., 2016) which will be discussed in Section 7.4.5 as a one-sweep method to determine $G_+$ for a fixed $G$.

In Bayesian clustering (rather than Bayesian mixture estimation), the main object of interest is the (marginal) posterior of the allocations $\mathbf{z}$, rather than the (marginal) posterior distribution of the mixture parameters $\theta$. Depending on the mixture under investigation, the integrated likelihood $p(\mathbf{y}|\mathbf{z}, G)$ for $G$ known may be available in closed form, in particular, if the component densities $f_g(y|\theta_g)$ come from exponential families and a conditionally conjugate prior $p(\theta_g)$ is employed for $\theta_g$. As noted, for instance, by Casella et al. (2004), for many mixture models it is then possible to derive an explicit form for the marginal posterior $p(\mathbf{z}|\mathbf{y}, G)$ of the indicators $\mathbf{z}$, where dependence on the parameter $\theta$ is integrated out. By Bayes' theorem, the marginal posterior $p(\mathbf{z}|\mathbf{y}, G)$ is given by

$$p(\mathbf{z}|\mathbf{y}, G) \propto p(\mathbf{y}|\mathbf{z}, G) p(\mathbf{z}|G), \tag{7.22}$$

where the integrated prior $p(\mathbf{z}|G)$ is given by (7.19) and the integrated likelihood $p(\mathbf{y}|\mathbf{z}, G)$ takes the form

$$p(\mathbf{y}|\mathbf{z}, G) = \int p(\mathbf{y}|\mathbf{z}, \theta_1, \ldots, \theta_G, \eta, G) p(\theta_1, \ldots, \theta_G, \eta|G) d(\theta_1, \ldots, \theta_G, \eta). \tag{7.23}$$

To explore the posterior of the allocations, efficient methods to sample from the posterior $p(\mathbf{z}|\mathbf{y}, G)$ are needed, and some of these methods will be discussed in Section 7.4.3. This exploration is quite a computational challenge, as the size of the lattice $\mathcal{S}_G^n$ increases rapidly with both the number $n$ of observations and the number $G$ of components and is given by the Bell number. For $n = 10$ and $G = 3$, for instance, there are 59,049 different allocations $\mathbf{z}$, whereas for $n = 100$ and $G = 3$ the number of different allocations is of the order of $5 \cdot 10^{47}$. This means that it is impossible to visit all possible partitions $\mathcal{C}$ during posterior sampling and many partitions are visited at best once.

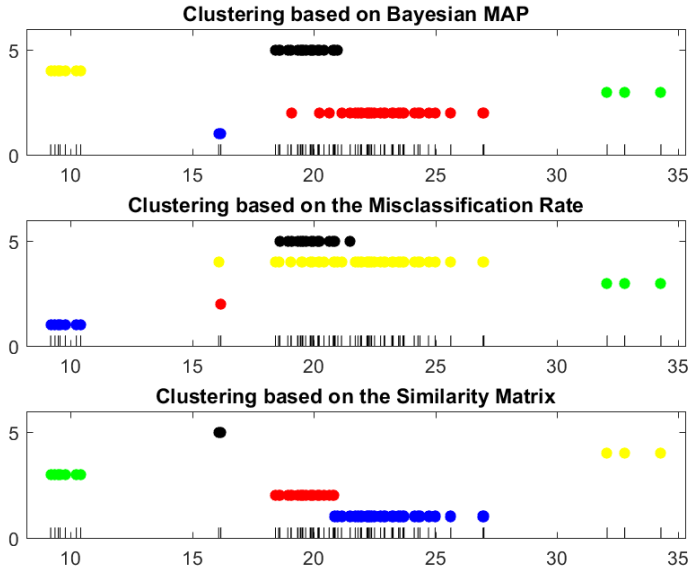


**FIGURE 7.5**
Bayesian clustering of the galaxy data (Roeder, 1990), assuming a Gaussian mixture with $G = 5$ components. The data are indicated through a rug plot. Partitions resulting from the MAP estimator (top), the minimum risk estimator (middle) and minimizing Binder's loss function.

This large set of partitions raises the question of how to summarize the posterior $p(\mathbf{z}|\mathbf{y}, G)$, given posterior simulations. Common summaries are based on deriving point estimators $\hat{\mathbf{z}}$, such as the MAP estimator, the minimum risk estimator or the partition minimizing Binder's loss function (Binder, 1978), see Section 8.3.2 for more details. However, these estimators (even if they differ) do not fully reflect the uncertainty in assigning observations to clusters.

By way of illustration, a mixture of univariate Gaussian distributions is used for Bayesian clustering of the galaxy data (Roeder, 1990), assuming that $G = 5$ is fixed. Prior specification follows Richardson & Green (1997), and 12,000 draws from $p(\mathbf{z}|\mathbf{y}, G)$ are obtained using full conditional Gibbs sampling. In Figure 7.5, various point estimators $\hat{\mathbf{z}}$ derived from the posterior draws of $\mathbf{z}$ are displayed, together with a rug plot of the data. While the MAP estimator and the estimator minimizing Binder's loss function are invariant to label switching, the minimum risk estimator is based on an identified model. Label switching is resolved by applying $k$-means clustering to the point process representation of the MCMC draws of $(\mu_g, \sigma_g)$. Classification over the various point estimators $\hat{\mathbf{z}}$ is stable for observations in the two clusters capturing the tails, but the classification for observations in the centre of the distribution tends to be rather different.

To quantify such uncertainty, Wade & Gharhamani (2018) develop not only appropriate point estimates, but also credible sets to summarize the posterior distribution of the partitions based on decision- and information-theoretic techniques.

### 7.3.4 Selecting $G$ under model misspecification

Mixture models are a very popular tool for model-based clustering, in both the frequentist and Bayesian frameworks. However, success in identifying meaningful clusters in the data very much hinges on specifying sensible component densities, and Bayesian inferences towards estimating the number of clusters are sensitive to misspecifications of the component densities, as are most penalized likelihood criteria discussed in the previous subsections. Most commonly, a finite mixture model with (multivariate) Gaussian component densities is fitted to the data to identify homogeneous data clusters within a heterogeneous population:

$$y \sim \sum_{g=1}^{G} \eta_g \mathcal{N}(\mu_g, \Sigma_g). \tag{7.24}$$

Similarly to the likelihood approach, Bayesian cluster analysis has to address several issues. First, as discussed above, even if we fit a correctly specified mixture model (7.1) to data generated by this model, an estimate of the number of components $G$ will not necessarily be a good estimator of the number of clusters $G_+$ in the data, and a more reliable estimate is obtained when exploring the partitions.

However, problems with the interpretation of $G_+$ might nevertheless occur, in particular if the component density is misspecified and several components have to be merged to address this misspecification. A typical example is fitting the multivariate Gaussian mixture distribution (7.24) to data such as the *Old Faithful Geyser* data. As shown in Figure 7.4, more than one Gaussian component is needed to capture departure from normality such as skewness and excess kurtosis for one of the two clusters. As discussed before, the BIC is particularly sensitive to this kind of misspecification, and classification-based information criteria such as the ICL criterion introduced in Section 7.3.2.1 are more robust in this respect.

In both Bayesian and frequentist frameworks, misspecification has been resolved by choosing more flexible distributions for the components densities. Many papers demonstrate the usefulness of mixtures of parametric non-Gaussian component densities in this context (see Frühwirth-Schnatter & Pyne, 2010, and Lee & McLachlan, 2013, among many others), and Chapter 10 also addresses this problem. Unsurprisingly, the estimated $G_+$ of such a non-Gaussian mixture often provides a much better estimator of the number of clusters than does the Gaussian mixture. With respect to inference, the Bayesian framework offers a slight advantage, as MCMC methods are able to deal with non-standard component densities in a more flexible way than the EM algorithm.

In higher dimensions it might be difficult to choose an appropriate parametric distribution for characterizing a data cluster, and mixture models with more flexible (not necessarily parametric) cluster densities turn out to be useful. The mixture of Gaussian mixtures approach, for instance, exploits the ability of normal mixtures to accurately approximate a wide class of probability distributions, and models the non-Gaussian cluster distributions themselves by Gaussian mixtures. This introduces a hierarchical framework where in the upper level a non-Gaussian mixture is fitted as in (7.1), whereas at a lower level each component density $f_g(y|\theta_g)$ itself is described by a mixture of $H_g$ Gaussian distributions. On the upper level, $G_+$ is identified as the number of such clusters, whereas the number of subcomponents $H_g$ in each cluster reflects the quality of the semi-parametric mixture approximation.

Two different approaches are available to "estimate" the number of clusters in such a framework. Any such approach has to deal with the following additional identifiability problems for this type of mixtures: the observed-data likelihood ascertains this model just as one big mixture of Gaussian distributions with $\tilde{G} = H_1 + \ldots + H_G$ components, and

it does not change when we exchange subcomponents between clusters on the lower level, even though this leads to different cluster distributions on the upper level of the mixture of mixtures model. Hence, a mixture of mixtures model is not identifiable in the absence of additional information, and this is most naturally dealt with within a Bayesian framework.

Within the Bayesian approach, it is common to estimate the hierarchical mixture of mixtures model directly by including such prior information; see, in particular, Malsiner-Walli et al. (2017) who consider a random-effects prior to introduce prior dependence among the $H_g$ means of the subcomponent Gaussian mixture defining $f_g(y|\theta_g)$. A different approach which is prevalent in the frequentist literature employs a two-step procedure and tries to create meaningful clusters after having fitted a Gaussian mixture as in (7.24) with $G = G_{\max}$. The clusters are determined by successively merging components according to some criterion such as the entropy of the resulting partition (Baudry et al., 2010); see Chapter 8, Section 8.2.2 for additional approaches and further details.

## 7.4 One-Sweep Methods for Cross-model Inference on $G$

From a Bayesian perspective, inference methods that treat $G$ or $G_+$ as an unknown parameter to be estimated jointly with the component-specific parameters $\theta$ are preferable to processing $G$ as a model index and relying on testing principles. Several such approaches are reviewed in this section.

### 7.4.1 Overfitting mixtures

Rousseau & Mengersen (2011) examine the issue of an overfitting mixture, that is, the estimation of a mixture model with $G$ components when the true distribution behind the data has fewer than $G$, say $G_0$, components. This setting complicates even further the non-identifiability of the mixture model, since there are $\binom{G}{G_0}$ ways of picking $G_0$ components out of the $G$ (while cancelling the others); see also Chapter 4, Section 4.2.2.

Rousseau & Mengersen (2011) show that the posterior distribution on the parameters of the overfitted mixture has a much more stable behaviour than the likelihood function when the prior on the weights of the mixture is sufficiently concentrated on the boundaries of the parameter space, that is, with many weights being close to zero. In fact, the central result of Rousseau & Mengersen (2011) is that, if the dimension $r$ of the component parameters is larger than twice the hyperparameter $e_0$ of a symmetric Dirichlet prior $\mathcal{D}_G(e_0)$ on the weights, then the sum of the weights of the extra $G - G_0$ components asymptotically concentrates at zero. This result has the additional appeal of validating less informative priors as asymptotically consistent. In practice, it means that selecting a Dirichlet $\mathcal{D}_G(1/2)$ and an arbitrary prior on the component parameters should see superfluous components vanish as the sample size grows to be large enough, even though the impact of the choice of $e_0$ can be perceived for finite sample sizes.

### 7.4.2 Reversible jump MCMC

Reversible jump MCMC (RJMCMC; Green, 1995) was exploited by Richardson & Green (1997) to select the number of components $G$ for univariate mixtures of Gaussian distributions. As briefly discussed in Chapter 1, Section 1.4.3, this simulation method is based on creating a Markov chain that moves over a space of variable dimensions, namely between the parameter spaces of finite mixtures with different numbers of components, while retaining

the fundamental detailed balance property that ensures the correct stationary (posterior) distribution.

The intuition behind the RJMCMC method is to create bijections between pairs of parameter spaces by creating auxiliary variables that equate the dimensions of the augmented spaces and to keep the same bijection for a move and its reverse. When designing a RJMCMC algorithm, those pairwise moves have to be carefully selected in order to reach sufficiently probable regions in the new parameter space. Richardson & Green (1997) discuss at length their *split-and-merge* moves which split (or aggregate) one (or two) components of the current mixture, with better performance than the basic *birth-and-death* moves, but performance may deteriorate as the number of components increases. The design of suitable proposals for higher-dimensional mixtures is quite a challenge, as demonstrated by Dellaportas & Papageorgiou (2006) and Zhang et al. (2004) for multivariate normal mixtures. In an attempt to extend RJMCMC methods to hidden Markov models, Cappé et al. (2002) had to face acceptance rates as low as 1%. RJMCMC is a natural extension of the traditional Metropolis–Hastings algorithm, but calibrating it is often perceived as too great an obstacle to its implementation, and it is not competitive with within-model simulations in the case of a small number of values of $G$ in competition.

### 7.4.3 Allocation sampling

As discussed in Section 7.3.3, the main object of interest in Bayesian clustering is the marginal posterior of the allocations, that is, $p(\mathbf{z}|\mathbf{y}, G)$ (if $G$ is known) or $p(\mathbf{z}|\mathbf{y})$ (if $G$ is unknown). Hence, Bayesian clustering has to rely on efficient methods to sample from the posterior $p(\mathbf{z}|\mathbf{y}, G)$ (or $p(\mathbf{z}|\mathbf{y})$).

While full conditional Gibbs sampling from the joint distribution $p(\theta, \mathbf{z}|\mathbf{y}, G)$ will yield draws from the (marginal) posterior $p(\mathbf{z}|\mathbf{y}, G)$, several authors considered alternative algorithms of "allocation sampling". Early Bayesian clustering approaches without parameter estimation are based on sampling from the marginal posterior distribution $p(\mathbf{z}|\mathbf{y}, G)$, defined earlier in (7.22), for known $G$. Chen & Liu (1996) were among the first to show how sampling of the allocation from $p(\mathbf{z}|\mathbf{y}, G)$ (for a fixed $G$) becomes feasible through MCMC methods, using either single-move Gibbs sampling or the Metropolis–Hastings algorithm; see Frühwirth-Schnatter (2006, Section 3.4) and Marin et al. (2005) for more details.

We want to stress here the following issue. Although these MCMC samplers operate in the marginal space of the allocations $\mathbf{z}$, neither the integrated likelihood $p(\mathbf{y}|\mathbf{z}, G)$, defined earlier in (7.23), nor the prior $p(\mathbf{z}|G)$, given in (7.19), can be (properly) defined without specifying a prior distribution $p(\theta_1, \ldots, \theta_G, \eta|G)$ for the unknown parameters of a mixture model with $G$ components. This problem is closely related to the problem discussed in Section 7.3.2.3 of having to choose priors for the exact ICL criterion. As discussed in Section 4.2.2, the choice of such a prior is not obvious and may have considerable impact on posterior inference.

These early sampling algorithms focus on computational aspects and do not explicitly account for the problem that the number $G_+$ of clusters in the sampled partitions $\mathbf{z}$ might differ from $G$, taking the identity of $G$ and $G_+$ more or less for granted. Still, as discussed above and again in Section 7.4.5, whether this applies or not very much depends on the choice of the hyperparameter $e_0$ in the Dirichlet prior $\mathcal{D}_G(e_0)$ on the weights.

Nobile & Fearnside (2007) address the problem of an unknown number of components $G$ in the context of allocation sampling. For a given $G$, they employ the usual Dirichlet prior $\eta|G \sim \mathcal{D}(e_1, \ldots, e_G)$ on the weight distribution, but treat $G$ as an unknown parameter, associated with a prior $p(G)$ (e.g. $G - 1 \sim \mathcal{P}(1)$), as justified by Nobile (2004). An MCMC sampler is developed that draws from the joint posterior $p(\mathbf{z}, G|\mathbf{y})$, by calling either Gibbs or Metropolis–Hastings moves based on the conditional distribution of $p(\mathbf{z}|\mathbf{y}, G)$ for a given

$G$ and by running RJMCMC type moves for switching values of $G$. Based on these posterior draws, $\hat{G}_+$ is estimated from the posterior draws of the number of non-empty clusters $G_+$. Several post-processing strategies are discussed for solving the label switching problem that is inherent in this sampler and for estimating $\hat{\mathbf{z}}$.

### 7.4.4 Bayesian nonparametric methods

A quite different approach of selecting the number $G_+$ of clusters exists outside the framework of finite mixture models and relies on Bayesian nonparametric approaches based on mixture models with countably infinite number of components, as discussed in Chapter 6 in full detail.

For Dirichlet process (DP) mixtures (Müller & Mitra, 2013), for instance, the discrete mixing distribution in the finite mixture (7.1) is substituted by a random distribution $H \sim DP(\alpha, H_0)$, drawn from a DP prior with precision parameter $\alpha$ and base measure $H_0$. As a draw $H$ from a DP is almost surely discrete, the corresponding model has a representation as an infinite mixture,

$$y \sim \sum_{g=1}^{\infty} \eta_g f_g(y|\theta_g), \tag{7.25}$$

with i.i.d. atoms $\theta_g \overset{iid}{\sim} H_0$ drawn from the base measure $H_0$ and weights $\eta_g$ obeying the stick-breaking representation

$$\eta_g = v_g \prod_{j=1}^{g-1} (1 - v_j), \quad g = 1, 2, \ldots, \tag{7.26}$$

with $v_g \sim \mathcal{B}e(1, \alpha)$ (Sethuraman, 1994).

As DP priors induce ties among the observations, such an approach automatically induces a random partition (or clustering) $\mathcal{C}$ of the data with a corresponding random cardinality $G_+$ (see Section 6.4). Since there are infinitely many components in (7.25) (i.e. $G = \infty$), there is no risk of confusing $G$ and $G_+$ as for finite mixtures. For a DP prior with precision parameter $\alpha$, the prior distribution over the partitions $\mathcal{C}$ is given by

$$p(\mathcal{C}|\alpha, G_+) = \alpha^{G_+} \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{g:n_g>0} \Gamma(n_g), \tag{7.27}$$

where $n_g$ and $G_+$ are defined as in (7.19). Another defining property of the DP prior is the prior predictive distribution $p(z_i|\mathbf{z}_{-i})$, where $\mathbf{z}_{-i}$ denotes all indicators excluding $z_i$. Let $G_+^{-i}$ be the number of non-empty clusters implied by $\mathbf{z}_{-i}$ and let $n_g^{-i}$, $g = 1, \ldots, G_+^{-i}$, be the corresponding cluster sizes. Then the probability that $z_i$ is assigned to an existing cluster $g$ is given by

$$\mathrm{P}(z_i = g|\mathbf{z}_{-i}, n_g^{-i} > 0) = \frac{n_g^{-i}}{n - 1 + \alpha}, \tag{7.28}$$

whereas the prior probability that $z_i$ creates a new cluster (indexed by $G_+^{-i} + 1$) is equal to

$$\mathrm{P}(z_i = G_+^{-i} + 1|\mathbf{z}_{-i}) = \frac{\alpha}{n - 1 + \alpha}. \tag{7.29}$$

Given this strong focus on BNP mixtures as random partition models, it is not surprising

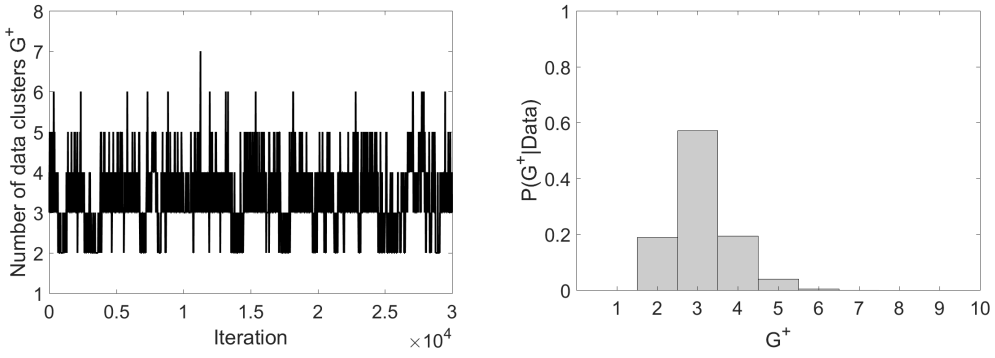


**FIGURE 7.6**
Sparse finite mixture modelling of the enzyme data: (left) 30,000 posterior draws of the number of data clusters $G_+$; (right) posterior distribution $p(G_+|\mathbf{y})$.

that the main interest in posterior inference is again in the draws from the posterior $p(\mathbf{z}|\mathbf{y})$ of the allocations which are exploited in various ways to choose an appropriate partition $\hat{\mathbf{z}}$ of the data and to estimate the number of clusters $G_+$.

Lau & Green (2007) compare BNP methods to estimate the number of clusters with the outcome associated with finite mixtures. They also show in detail how to derive a single (optimal) point estimate $\hat{\mathbf{z}}$ from the posterior $p(\mathbf{z}|\mathbf{y})$, with the number of distinct clusters $\hat{G}_+$ in $\hat{\mathbf{z}}$ being an estimator of $G_+$ in this framework. To derive a partition of the data, Molitor et al. (2010) cluster the data using the pairwise association matrix as a distance measure which is obtained by aggregating over all partitions obtained during MCMC sampling, using partitioning around medoids. The optimal number of clusters is determined by maximizing an associated clustering score; see also Liverani et al. (2013).

A well-known limitation of DP priors is that *a priori* the cluster sizes are expected to be geometrically ordered, with one big cluster, geometrically smaller clusters, and many singleton clusters (Müller & Mitra, 2013). This initiated the investigation of alternative BNP mixtures and their usefulness for clustering. A popular BNP two-parameter mixture is obtained from the Pitman–Yor process (PYP) prior $PY(\beta, \alpha)$ with $\beta \in [0, 1)$, $\alpha > -\beta$ (Pitman & Yor, 1997), with a stick-breaking representation as in (7.26) with $v_g \sim \mathcal{B}e(1 - \beta, \alpha + k\beta)$. The DP prior occurs as a special case when $\beta = 0$. PYP mixtures are known to be more useful than the DP mixture for data with many significant, but small clusters.

For a DP as well as a PYP mixture, the prior expected number of data clusters $G_+$ increases as the number $n$ of observations increases, where for the DP process $G_+ \sim \alpha \log(n)$ (Korwar & Hollander, 1973) and $G_+ \sim n^\beta$ obeys a power law for PYP mixtures. As will be discussed in the next subsection, finite mixtures are quite different in this respect.

### 7.4.5 Sparse finite mixtures for model-based clustering

Inspired by the important insights of Rousseau & Mengersen (2011), Malsiner-Walli et al. (2016) introduced the concept of sparse finite mixture models for model-based clustering as an alternative to infinite mixtures, following ideas presented earlier in Frühwirth-Schnatter (2011). A similar approach is pursued by van Havre et al. (2015).

While remaining within the framework of finite mixtures, sparse finite mixture models provide a semi-parametric Bayesian approach in so far as the number $G_+$ of non-empty

mixture components used to generate the data is not assumed to be known in advance, but random, as already discussed in Section 7.3.3. The basic idea of sparse finite mixture modelling is to deliberately specify an *overfitting* finite mixture model with too many components $G$. Sparse finite mixtures stay within the common finite mixture framework by assuming a symmetric Dirichlet prior $\eta \sim \mathcal{D}_G(e_0)$ on the weight distribution; however, the hyperparameter $e_0$ of this prior is selected such that superfluous components are emptied automatically during MCMC sampling and sparse solutions with regard to the number $G_+$ of clusters are induced through the prior on the weight distribution. This proposal leads to a simple Bayesian framework where a straightforward MCMC sampling procedure is applied to jointly estimate the unknown number of non-empty data clusters $G_+$ with the remaining parameters.

As discussed in Section 7.3.3, for such a mixture model, the number $G$ of components does not reflect the number of data clusters, as many components will remain unused. Following Nobile (2004), Malsiner-Walli et al. (2016) derive the posterior distribution $\mathrm{P}(G_+ = g|\mathbf{y}),\ g = 1, \ldots, G$, of the number $G_+$ of data clusters from the MCMC output of the allocations $\mathbf{z}$. Therefore, for each iteration $m$ of MCMC sampling, all components $g$ to which some observations have been assigned are identified from $\mathbf{z}^{(m)}$ and the corresponding number of non-empty components is considered:

$$G_+^{(m)} = G - \sum_{g=1}^{G} \mathbb{I}(n_g^{(m)} = 0),$$

where, for $g = 1, \ldots, G$, $n_g^{(m)} = \sum_{i=1}^{n} \mathbb{I}(z_i^{(m)} = g)$ is the number of observations allocated to component $g$, and $\mathbb{I}(\cdot)$ denotes the indicator function. The posterior distribution $\mathrm{P}(G_+ = g|\mathbf{y})$, $g = 1, \ldots, G$, is then estimated by the corresponding relative frequency.

The number of clusters $G_+$ can be derived as a point estimator from this distribution, for example, the posterior mode estimator $\tilde{G}_+$ that maximizes the (estimated) posterior distribution $\mathrm{P}(G_+ = g|\mathbf{y})$. This happens to be the most frequent number of clusters visited during MCMC sampling. The posterior mode estimator appears to be sensible in the present context when adding very small clusters hardly changes the marginal likelihood. This makes the posterior distribution $\mathrm{P}(G_+ = g|\mathbf{y})$ extremely right-skewed, and other point estimators such as the posterior mean are extremely sensitive to prior choices, as noted by Nobile (2004). However, under a framework where sparse finite mixtures are employed for density estimation, very small components might be important and other estimators of $G_+$ might be better justified.

An alternative way to summarize clustering based on sparse finite mixtures is by exploring the posterior draws of the partitions $\mathbf{z}$ and determining some optimal partition, such as the partition $\hat{\mathbf{z}}$ minimizing Binder's loss function. This can be done without the need to resolve label switching or to stratify the draws with respect to $G_+$. The cardinality $\hat{G}_+$ of such an optimal partition $\hat{\mathbf{z}}$ is yet another estimator of the number of clusters. The posterior mode estimator $\tilde{G}_+$ and $\hat{G}_+$ do not necessarily coincide, and differences in these estimators reflect uncertainty in the posterior distribution over the partition space. As discussed by Frühwirth-Schnatter et al. (2018), the approach of Wade & Gharhamani (2018) to quantifying such uncertainty can be applied immediately to sparse finite mixture models.

The appropriate choice of the hyperparameter $e_0$ is important for the application of the sparse finite mixture approach in a clustering context. While in a density estimation framework the asymptotic criterion of Rousseau & Mengersen (2011) suggests the choice $e_0 < r/2$, with $r$ being the dimension of $\theta_g$, this rule is not necessarily a sensible choice for selecting the number of clusters $G_+$ in a data set of finite size $n$, as demonstrated for a broad range of mixture models in Malsiner-Walli et al. (2016) and Frühwirth-Schnatter

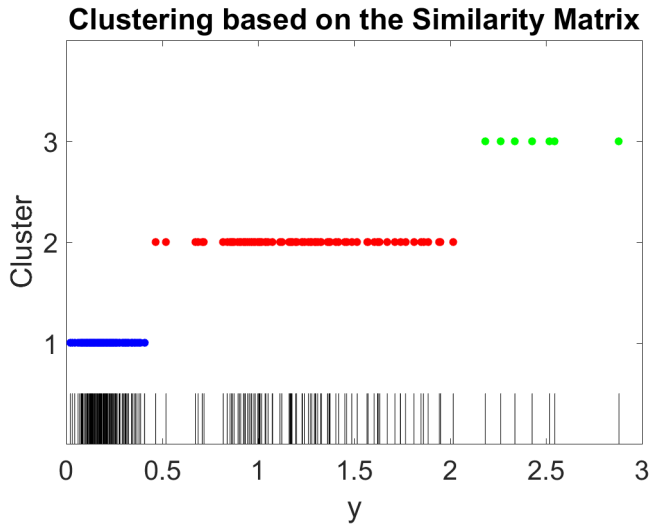


**FIGURE 7.7**
Sparse finite mixture modelling of the enzyme data, displayed as a rug plot. Partition $\hat{\mathbf{z}}$ optimizing Binder's loss function. The number of clusters in this partition is equal to three.

& Malsiner-Walli (2018). Indeed, these papers show that values of $e_0 \ll r/2$ much smaller than the asymptotic criterion of Rousseau & Mengersen (2011) are needed to identify the right number of clusters, and recommend choosing either very small fixed values such as $e_0 = 0.001$ or applying a hyperprior with $e_0 \sim \mathcal{G}(a_e, b_e)$ such that $\mathrm{E}(e_0) = a_e/b_e$ is very small (e.g. $e_0 \sim \mathcal{G}(1, 200)$).

Under the provision that $G_+$ underestimates $G$, this approach constitutes a simple and generic strategy for model selection without making use of model selection criteria, RJMCMC, or marginal likelihoods. Applications include Gaussian mixtures as well as mixtures of Gaussian mixtures (Malsiner-Walli et al., 2017) and sparse mixtures for discrete-valued data (Frühwirth-Schnatter & Malsiner-Walli, 2018). By way of further illustration, the enzyme data (shown earlier in Figure 7.2) are reanalysed using sparse finite mixtures, taking the prior of Richardson & Green (1997) as base measure. The maximum number of data clusters is chosen as $G = 10$ and the hierarchical sparse Dirichlet prior $\eta \sim \mathcal{D}_G(e_0)$, $e_0 \sim \mathcal{G}(1, 200)$ is applied.

Figure 7.6 shows 30,000 posterior draws of the number of data clusters $G_+$ as well as the corresponding posterior distribution $p(G_+|\mathbf{y})$. The posterior mode estimator yields three clusters with $\mathrm{P}(G_+ = 3|\mathbf{y}) = 0.57$. Also two clusters are supported with $\mathrm{P}(G_+ = 2|\mathbf{y}) = 0.19$, which is not unexpected in the light of Figure 7.2, showing two (albeit non-Gaussian) data clusters. Due to this misspecification of the component densities the four-cluster solution is equally supported with $\mathrm{P}(G_+ = 4|\mathbf{y}) = 0.19$. Finally, Figure 7.7 shows the partition $\hat{\mathbf{z}}$ optimizing Binder's loss function together with a rug plot of the data. The number of clusters in this partition is equal to three, supporting the choice based on the posterior mode. The resulting clustering nicely captures the three distinct groups of data points.

*Relation to BNP methods*

The concept of sparse finite mixtures is related in various ways to DP mixtures, discussed in Section 7.4.4. If the weight distribution follows the Dirichlet prior $\eta \sim \mathcal{D}_G(\alpha/G)$ and the base measure $H_0$ serves as prior for the component parameters (i.e. $\theta_g \sim H_0$), then, as shown by Green & Richardson (2001), the finite mixture in (7.1) converges to a DP mixture with mixing distribution $H \sim DP(\alpha, H_0)$ as $G$ increases. This relationship has mainly been exploited to obtain a finite mixture approximation to the DP mixture. In this sense, the sparse finite Gaussian mixture introduced in Malsiner-Walli et al. (2016) could be seen as an approximation to a DP mixture. Nevertheless, as argued by Malsiner-Walli et al. (2017), it makes sense to stay within the framework of finite mixtures and to consider $G$ as a second parameter which is held fixed at a finite value, as this provides a two-parameter alternative to DP mixtures with related properties.

Representations similar to BNP mixtures exist also for finite mixture models under the symmetric prior $\eta \sim \mathcal{D}_G(e_0)$, but are not commonly known, although they shed further light on the relation between the two model classes. First of all, a stick-breaking representation of the weights $\eta_1, \eta_2, \ldots, \eta_G$ as in (7.26) in terms of a sequence of independently (albeit not identically) distributed random variables exists also for finite mixtures, with $v_g \sim \mathcal{B}e(e_0, (G-g)e_0)$, $g = 1, \ldots, G-1$, $v_G = 1$; see, for example, Frühwirth-Schnatter (2011).

Second, as already discussed in Section 7.3.1, finite mixture models can be regarded as random partition models and the prior distribution over all random partitions $\mathcal{C}$ of $n$ observations can be derived from the joint (marginal) prior $p(\mathbf{z}|G)$ given in (7.19) (see, for example, Malsiner-Walli et al. (2017)):

$$p(\mathcal{C}|e_0, G_+) = \frac{G!}{(G-G_+)!} \frac{\Gamma(Ge_0)}{\Gamma(n+Ge_0)\Gamma(e_0)^{G_+}} \prod_{g:n_g>0} \Gamma(n_g + e_0).$$

takes the form of a product partition model as for DP mixtures (see (7.27)) and is invariant to permuting the cluster labels.

Finally, as for BNP mixtures, it is possible to derive the prior predictive distribution $p(z_i|\mathbf{z}_{-i})$, where $\mathbf{z}_{-i}$ denotes all indicators excluding $z_i$. Let $G_+^{-i}$ be the number of non-empty clusters implied by $\mathbf{z}_{-i}$, and let $n_g^{-i}$, $g = 1, \ldots, G_+^{-i}$, be the corresponding cluster sizes. Then the probability that $z_i$ is assigned to an existing cluster $g$ is given by

$$\mathrm{P}(z_i = g|\mathbf{z}_{-i}, n_g^{-i} > 0) = \frac{n_g^{-i} + e_0}{n - 1 + e_0 G},$$

which is closely related to (7.28), in particular if $e_0 = \alpha/G$ and $G$ increases. However, the prior probability that $z_i$ creates a new cluster with $z_i \in I = \{g|n_g^{-i} = 0\}$ is equal to

$$\mathrm{P}(z_i \in I|\mathbf{z}_{-i}) = \frac{e_0(G - G_+^{-i})}{n - 1 + e_0 G}, \tag{7.30}$$

and is quite different from (7.29). In particular, for $e_0$ independent of $G$, this probability not only depends on $e_0$, but also increases with $G$. Hence a sparse finite mixture model can be regarded as a two-parameter model, where both $e_0$ and $G$ influence the prior expected number of data clusters $G_+$, which is determined for a DP mixture solely by $\alpha$. Furthermore, the prior probability (7.30) of creating new clusters decreases as the number $G_+^{-i}$ of non-empty clusters increases, as opposed to DP mixtures where this probability is constant and to PYP mixtures where this probability increases. Hence, sparse finite mixtures are useful for clustering data that arise from a moderate number of clusters that does not increase as the number of data points $n$ increases.

## 7.5 Concluding Remarks

The issue of selecting the number of mixture components has always been contentious, both in frequentist and Bayesian terms, and this chapter has reflected on this issue by presenting a wide variety of solutions and analyses. The main reason for the difficulty in estimating the order $G$ of a mixture model is that it is a poorly defined quantity, even when setting aside identifiability and label switching aspects. Indeed, when considering a single sample of size $n$ truly generated from a finite mixture model, there is always a positive probability that the observations in that sample are generated from a subset of the components of the mixture of size $G_+$ rather than from all components $G$. As shown by the asymptotic results in Chapter 4, the issue goes away as the sample size $n$ goes to infinity (provided $G$ remains fixed), but this does not bring a resolution to the quandary of whether or not $G$ is estimable. In our opinion, inference should primarily bear on the number of data clusters $G_+$, since the conditional posterior distribution of $G$ given $G_+$ mostly depends on the prior modelling and very little on the data. Without concluding like Larry Wasserman (on his now defunct *Normal Deviate* blog) that "mixtures, like tequila, are inherently evil and should be avoided at all costs", we must acknowledge that the multifaceted uses of mixture models imply that the estimation of a quantity such as the number of mixture components should be impacted by the purpose of modelling via finite mixtures, as for instance through the prior distributions in a Bayesian setting.

# Bibliography


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Banfield, J. D. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Baudry, J.-P. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics* **9**, 1041–1077.

Baudry, J.-P., Maugis, C. & Michel, B. (2012). Slope heuristics: Overview and implementation. *Statistics and Computing* **22**, 455–470.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K. & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* **19**, 332–353.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 2nd ed.

Berger, J. & Jefferys, W. (1992). Sharpening Ockham's razor on a Bayesian strop. *American Statistician* **80**, 64–72.

Berkhof, J., van Mechelen, I. & Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica* **13**, 423–442.

Bertoletti, M., Friel, N. & Rastelli, R. (2015). Choosing the number of components in a finite mixture using an exact integrated completed likelihood criterion. *Metron* **73**, 177–199.

Besag, J. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika* **76**, 183.

Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719–725.

Biernacki, C., Celeux, G. & Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning Inference* **140**, 2991–3002.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Birgé, L. & Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3**, 203–268.

Birgé, L. & Massart, P. (2007). Gausian model selection. *Probability Theory and Related Fields* **138**, 33–73.

Bouveyron, C., Côme, E. & Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Annals of Applied Statistics* **9**, 1726–1760.

Box, G. E. P. (1976). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association* **71**, 791–799.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

Cappé, O., Robert, C. P. & Rydén, T. (2002). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *Journal of the Royal Statistical Society, Series B* **65**, 679–700.

Casella, G., Robert, C. P. & Wells, M. T. (2004). Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology* **1**, 1–18.

Celeux, G., Chrétien, S., Forbes, F. & Mkhadri, A. (2001). A component-wise EM algorithm for mixtures. *Journal of Computational and Graphical Statistics* **10**, 697–712.

Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–674.

Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification* **13**, 195–212.

Chen, R. & Liu, J. S. (1996). Predictive updating methods with application to Bayesian classification. *Journal of the Royal Statistical Society, Series B* **58**, 397–415.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika* **89**, 539–552.

Chopin, N. & Robert, C. P. (2010). Properties of nested sampling. *Biometrika* **97**, 741–755.

Côme, E. & Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling* **6**, 564–589.

Dacunha-Castelle, D. & Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3**, 279–299.

Dacunha-Castelle, D. & Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Annals of Statistics* **27**, 1178–1209.

De Iorio, M. & Robert, C. P. (2002). Discussion of "Bayesian measures of model complexity and fit", by Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B* **64**, 629–630.

Dellaportas, P. & Papageorgiou, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing* **16**, 57–68.

DiCiccio, T. J., Kass, R. E., Raftery, A. & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.

Drton, M. & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society, Series B* **79**, 1–18.

Figueiredo, M. & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 381–396.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Frühwirth-Schnatter, S. (1995). Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of the Royal Statistical Society, Series B* **57**, 237–246.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96**, 194–209.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal* **7**, 143–167.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag.

Frühwirth-Schnatter, S. (2011). Dealing with label switching under model uncertainty. In *Mixture Estimation and Applications*, K. Mengersen, C. P. Robert & D. Titterington, eds., chap. 10. Chichester: Wiley, pp. 193–218.

Frühwirth-Schnatter, S. (2018). *Applied Bayesian Mixture Modelling. Implementations in MATLAB using the package bayesf Version 4.0.* https://www.wu.ac.at/statmath/faculty-staff/faculty/sfruehwirthschnatter/.

Frühwirth-Schnatter, S., Grün, B. & Malsiner-Walli, G. (2018). Contributed comment on article by Wade and Gharamani. *Bayesian Analysis* **13**, 601–603.

Frühwirth-Schnatter, S. & Malsiner-Walli, G. (2018). From here to infinity – sparse finite versus Dirichlet process mixtures in model-based clustering. Preprint, arXiv:1706.07194.

Frühwirth-Schnatter, S. & Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew normal and skew-$t$ distributions. *Biostatistics* **11**, 317–336.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques* **38**, 887–906.

Gassiat, E. & van Handel, R. (2013). Consistent order estimation and minimal penalties. *IEEE Transactions on Information Theory* **59**, 1115–1128.

Gelfand, A. & Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gelfand, A. & Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. & Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P. J. & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Hansen, B. (1992). The likelihood ratio test under non-standard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics* **7**, S61–S82.

Keribin, C. (2002). Consistent estimation of the order of mixture models. *Sankhyā, Series A* **62**, 49–66.

Keribin, C., Brault, V., Celeux, G. & Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing* **25**, 1201–1216.

Korwar, R. M. & Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability* **1**, 705–711.

Lange, K. (1999). *Numerical Analysis for Statisticians*. New York: Springer-Verlag.

Lau, J. W. & Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

Lee, J. & Robert, C. (2016). Importance sampling schemes for evidence approximation in mixture models. *Bayesian Analysis* **11**, 573–597.

Lee, K., Marin, J.-M., Mengersen, K. & Robert, C. (2009). Bayesian inference on mixtures of distributions. In *Perspectives in Mathematical Sciences I: Probability and Statistics*, N. N. Sastry, M. Delampady & B. Rajeev, eds. Singapore: World Scientific, pp. 165–202.

Lee, S. X. & McLachlan, G. J. (2013). EMMIXuskew: An R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm. *Journal of Statistical Software* **55**, 1–22.

Liverani, S., Hastie, D. I., Papathomas, M. & Richardson, S. (2013). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. Preprint, arXiv:1303.2836.

Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing* **26**, 303–324.

Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics* **26**, 285–295.

Marin, J.-M., Mengersen, K. & Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics*, C. R. Rao & D. Dey, eds., vol. 25. New York: Springer-Verlag, pp. 459–507.

McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Meng, X. & Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Miller, J. W. & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**, 340–356.

Molitor, J., Papathomas, M., Jerrett, M. & Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics* **11**, 484–498.

Müller, P. & Mitra, R. (2013). Bayesian nonparametric inference – why and how. *Bayesian Analysis* **8**, 269–360.

Nadif, M. & Govaert, G. (1998). Clustering for binary data and mixture models – choice of the model. *Applied Stochastic Models and Data Analysis* **13**, 269–278.

Neal, R. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". Tech. rep., University of Toronto.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics* **32**, 2044–2073.

Nobile, A. & Fearnside, A. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing* **17**, 147–162.

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25**, 855–900.

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L. & Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics* **31**, 1420–1427.

Richardson, S. & Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Rissanen, J. (2012). *Optimal Estimation of Parameters*. Cambrdige: Cambridge University Press.

Robert, C. P. (2007). *The Bayesian Choice*. New York: Springer.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2nd ed.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association* **85**, 617–624.

Roeder, K. & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894–902.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society, Series B* **73**, 689–710.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4**, 10–26.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Skilling, J. (2007). Nested sampling for Bayesian computations. *Bayesian Analysis* **1**, 833–859.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

Tessier, D., Schoenauer, M., Biernacki, C., Celeux, G. & Govaert, G. (2006). Evolutionary latent class clustering of qualitative data. Tech. Rep. RR-6082, INRIA.

van Havre, Z., White, N., Rousseau, J. & Mengersen, K. (2015). Overfitting Bayesian mixture models with an unknown number of components. *PLoS ONE* **10**, e0131739.

Wade, S. & Gharhamani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis* **13**, 559–626.

Wallace, C. & Freeman, P. (1987). Estimation and inference via compact coding. *Journal of the Royal Statistical Society, Series B* **49**, 241–252.

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge: Cambridge University Press.

Wyse, J., Friel, N. & Latouche, P. (2017). Inferring structure in bipartite networks using the latent block model and exact ICL. *Network Science* **5**, 45–69.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937–950.

Zeng, H. & Cheung, Y.-M. (2014). Learning a mixture model for clustering with the completed likelihood minimum message length criterion. *Pattern Recognition* **47**, 2011–2030.

Zhang, Z., Chan, K. L., Wu, Y. & Chen, C. (2004). Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. *Statistics and Computing* **14**, 343–355.